



National Institute of Environmental Health Sciences
Your Environment. Your Health.

DNA Microarray Analysis

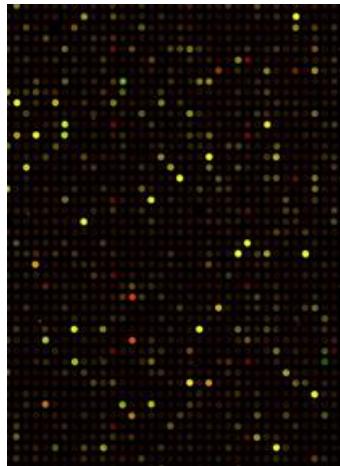
**Keith R. Shockley
NIEHS BB**

June 18, 2013

Outline

- Microarray Technology
- Experimental Design
- Quality Control
- Data Preprocessing
- Testing for Differential Expression (ANOVA)
- Clustering

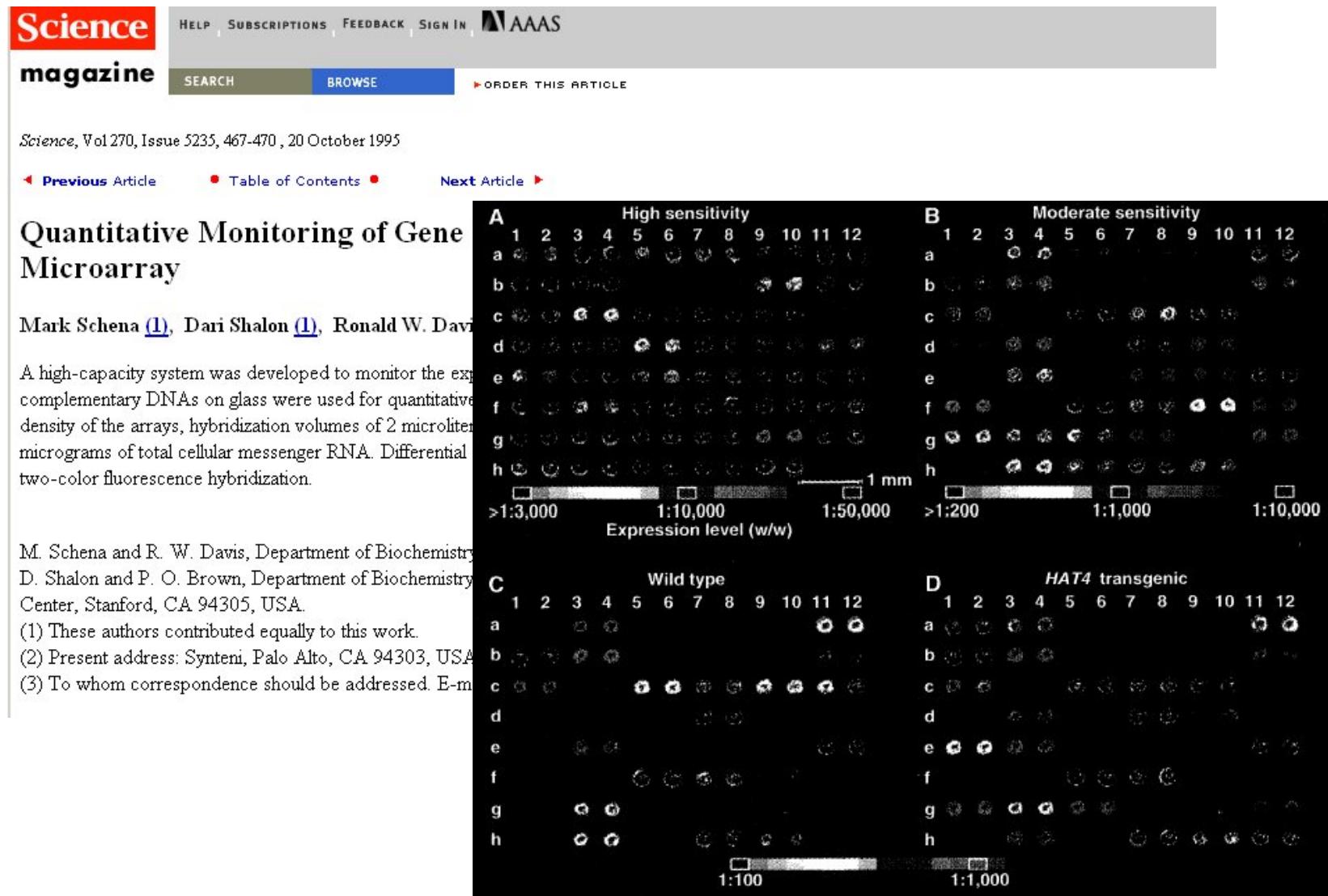
DNA Microarrays



A hybridization-based platform able to profile gene expression patterns of tens of thousands of genes in a single experiment to gain information about the state of a particular tissue sample at a particular time under a particular set of conditions

image source: http://www.frontiers-in-genetics.org/pictures/genechip_1.jpg

The First DNA Microarray Paper



Types of Microarrays

Gene expression profiling

Comparative genomic hybridization

Tissue (TMAs)

miRNA (microRNA profiling)

Methylation analysis

Resequencing assays

Genotyping (SNPs)

Proteome profiling analysis

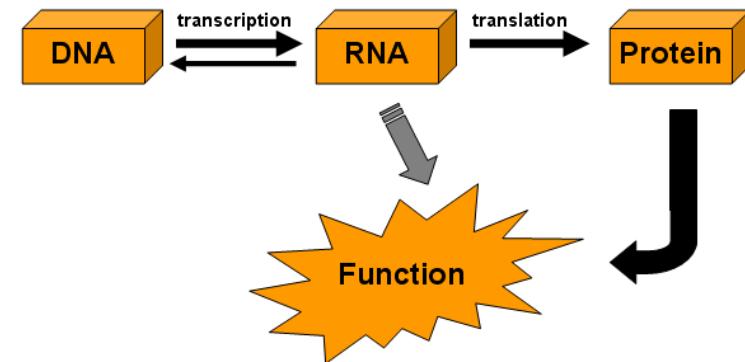
Cellular (proteins, lipids, antibodies)

Protein phosphorylation & glycosylation analysis

Chemical compound interrogation

Chromatin immunoprecipitation (ChIP)

Protein: DNA interaction analysis



Gene Expression Microarrays

- Spotted arrays
(PCR, oligo, cDNA)
- Affymetrix GeneChip® platform
- Commercial spotted/printed/in situ oligo arrays
(Agilent, Compugen, Qiagen, Nimblegen)
- Bead arrays
(Illumina and Luminex)
- Other

Raw Data: NOT mRNA concentrations!

- RNA Quality (e.g., degradation, purity)
- Amplification Efficiency
- Reverse Transcription Efficiency
- Hybridization Efficiency and Specificity
- Clone Identification and Mapping
- PCR yield and Contamination
- Manufacturing Issues (e.g., batch effects)
- Signal Quantification (e.g., faulty scanner)
- Background Correction
- Tissue Contamination

see also: http://www.cs.princeton.edu/picasso/mats/microarray_analysis_basics_F08

Need adequate experimental design and analysis

Microarray Analysis Workflow



Experimental design: define question/goals
include replication

Quality control: remove outlier probes or arrays
are samples sufficiently similar?

Data preprocessing: remove poor quality or control spots
 \log_2 transform data

Data normalization: correct for background effects
balance signal intensities between arrays
summarize any “probe sets”

Statistical testing: test for differential expression
apply multiple hypothesis test correction
select significance threshold to find gene list

Clustering: find genes (conditions) with similar expression profiles

Bioinformatic analysis: study gene list in context of known biology
look for enrichment of structural features
compare experimental results to literature

R and Bioconductor

R and R package systems are used to design and distribute software. Most bioconductor components are distributed as R packages.

“Bioconductor is an open source and open development software project to provide tools for the analysis and comprehension of genomic data (bioinformatics).”

source: <http://www.bioconductor.org>

Example R script: (example from R/affy package to normalize microarray data)

```
library(affyPLM);
ReadData = Readaffy();
norm = rma();
RMAexpr = exprs(norm);
Processednorm = data.frame(cloneid=row.names(RMAexpr),RMAexpr);
write.table(Processednorm,"rma.dat",sep="\t",row=F,quote=F);
design.table=data.frame(Array = row.names(pData(ReadData)));
write.table(design.table,"design.dat",sep="\t",row=F,quote=F);
```

Experimental Design

Experimental Design

“A well-designed experiment will usually allow its conclusions to be easily obtained, whereas no computations, however industriously or ingeniously performed, can produce entirely satisfactory conclusions from an ill-designed one.”

-Finney, 1953

Experimental Design

- The set of treatments selected for comparison
- The specification of the units (mouse, cage, pool) to which the treatments will be applied
- The rules by which the treatments are allocated to experimental units
- The specifications of the measurements to be made

Experimental Design

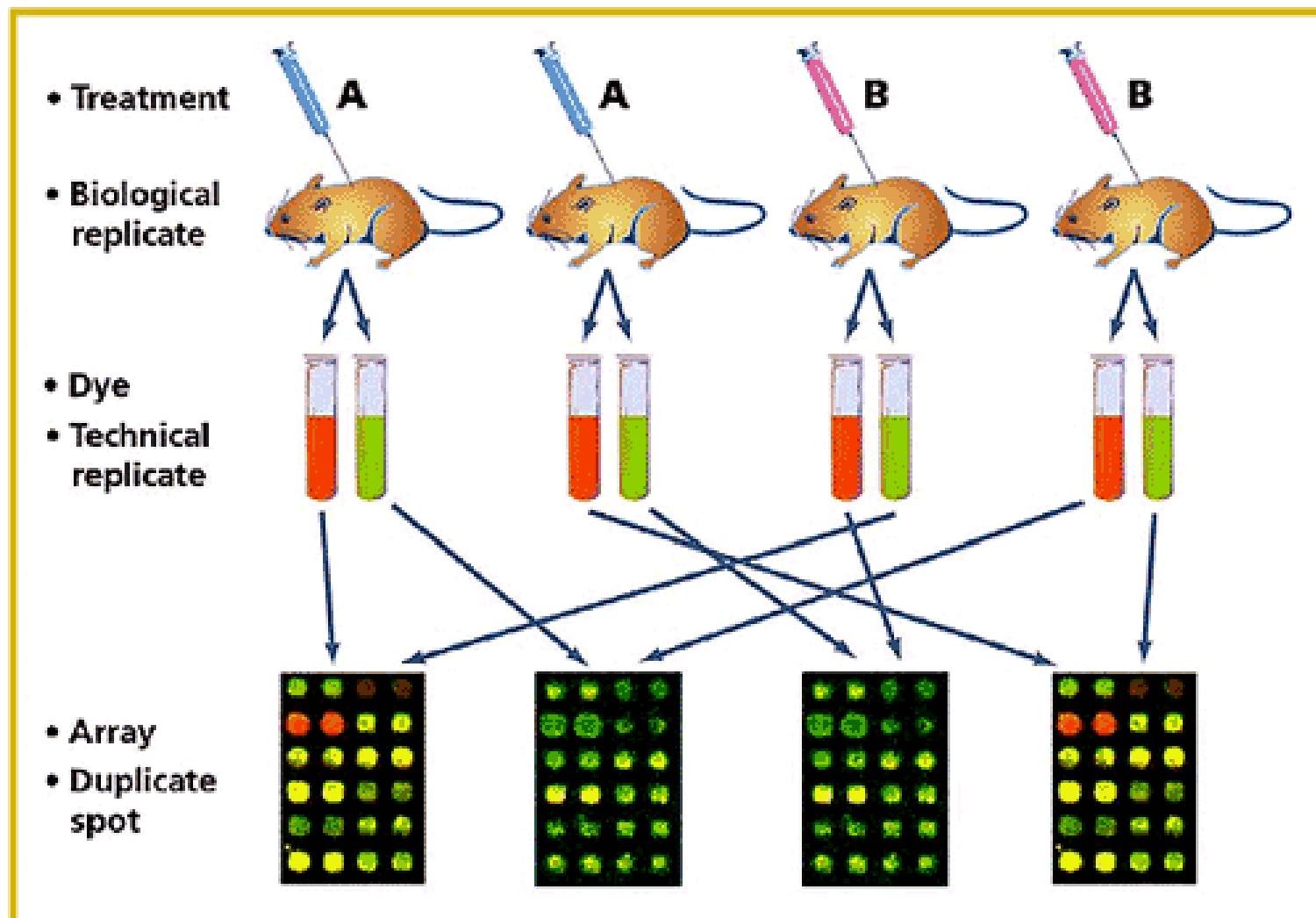
Sources of Variation

- Treatment
 - Drug
 - Tissue
 - Strain
- Technical variation
 - RNA extraction
 - Labeling
 - Hybridization
 - Systems/devices
- Random noise

Replication

- Experimental unit
 - Independently received Trt
 - person, mouse, fish, etc
- Biological replicates
 - True replicates
 - Experimental unit
- Technical replicates
 - “pseudo-replicates”

Experimental Design



Churchill, *Nature Genetics*, 2002

Sources of Variation

- A. Differences due to treatments
- B. Intrinsic biological variation (mouse)
- C. Technical variation in extraction and labeling of RNA samples
- D. Technical variation in hybridization
- E. Spot size variation (e.g., spotted arrays)
- F. Measurement error in scanning

Resource Allocation

- Replicated Spots Reduce E, F
(quality control)
- Multiple Arrays per Sample Reduce D, E, F
(estimate technical variance)
- Multiple Samples per Group Reduce B, C, D, E, F
(estimate biological variance)
- Pooling Reduce B

source: pga.jax.org/hlb06coursefiles/Churchill_10-26.pdf

$$EV \approx \frac{\sigma_M^2}{m} + \frac{\sigma_e^2}{mn}$$

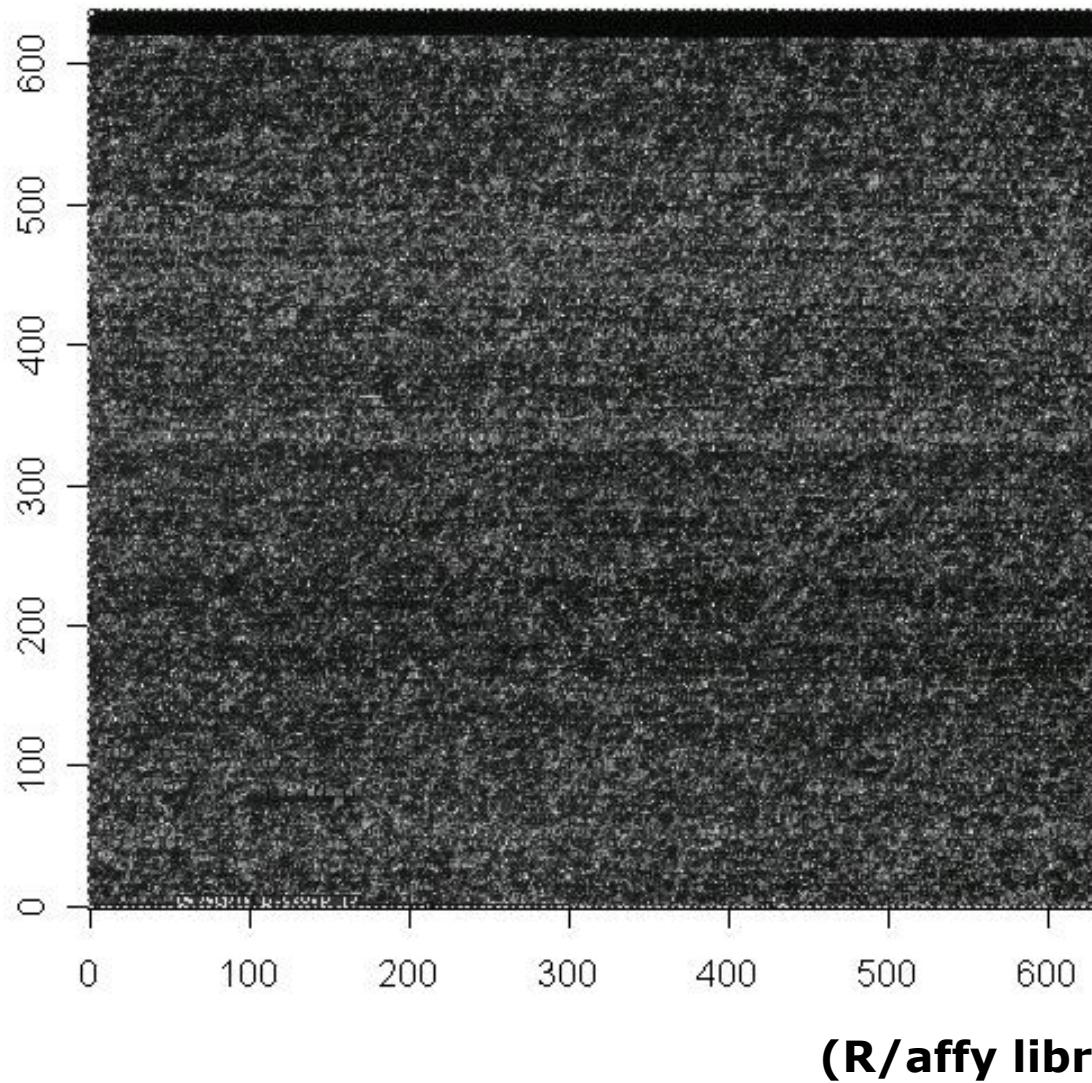
m: number of mice per treatment
n: number of array pairs per mouse

Experimental Design

- Use Independent Biological Replicates
(use as many as possible!)
- Stay focused on the goal of the study
- Randomize as much as possible!
 - Treatment assignments
 - sample populations
 - spot locations
 - etc

Quality Control

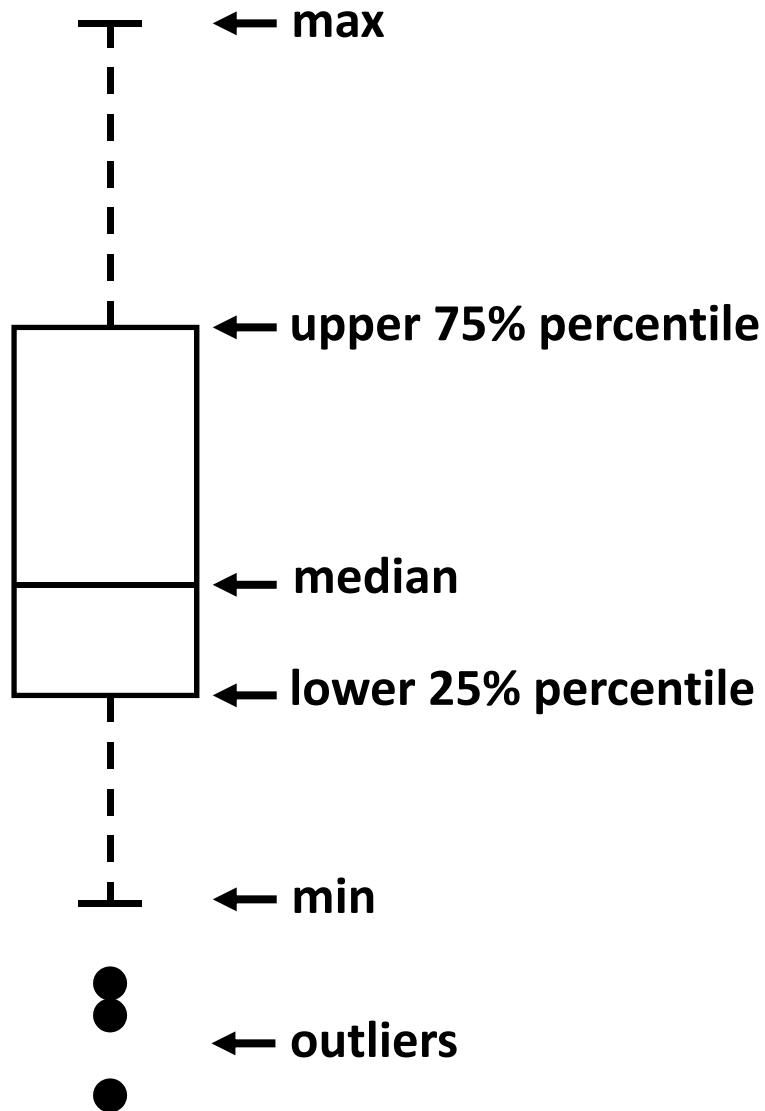
QC: Reconstruct Images



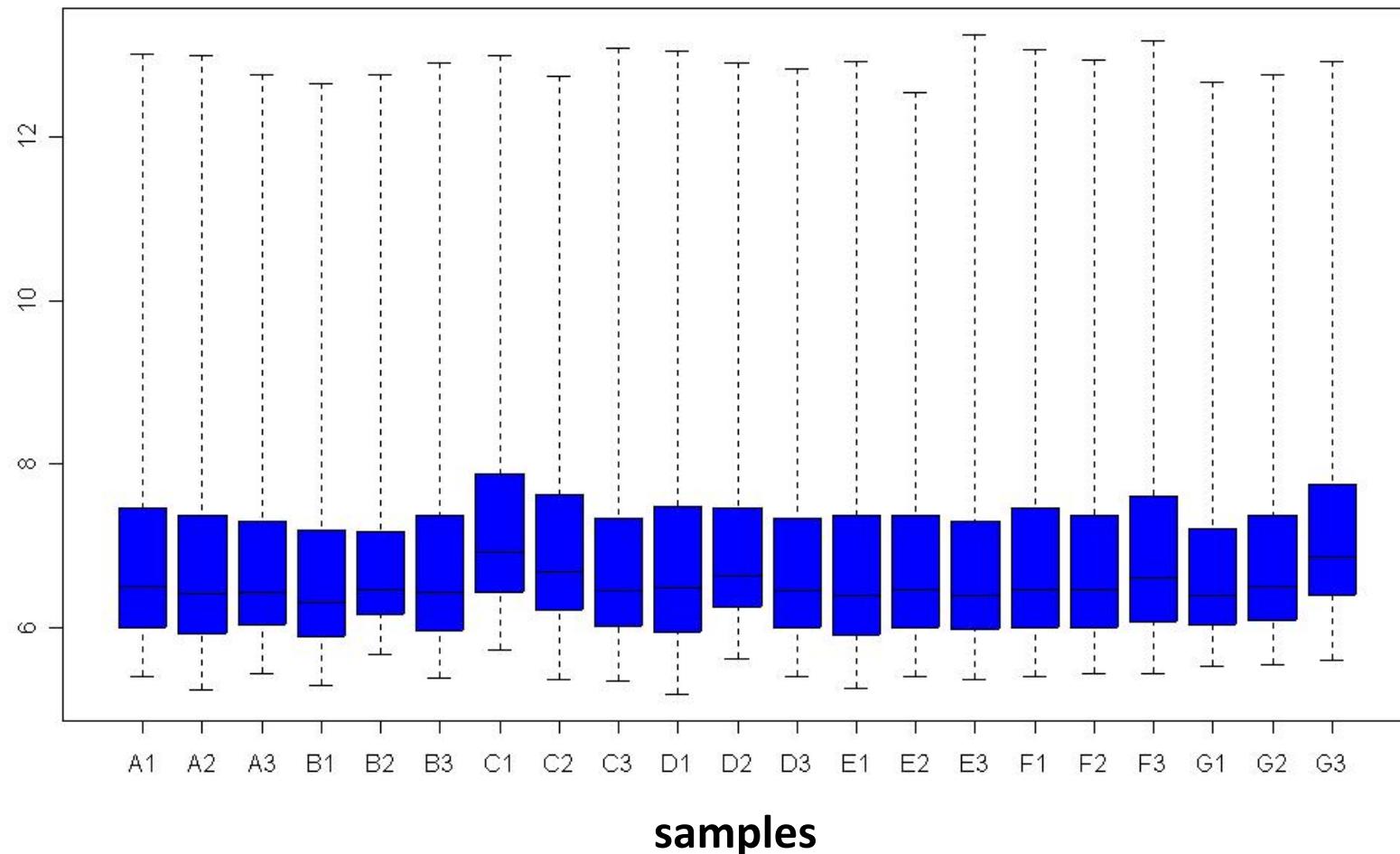
Visually Inspect:

- Dimness/Brightness
- Background
- Scratches/Cracks
- Unevenness
- Banding
- Spots

Boxplots

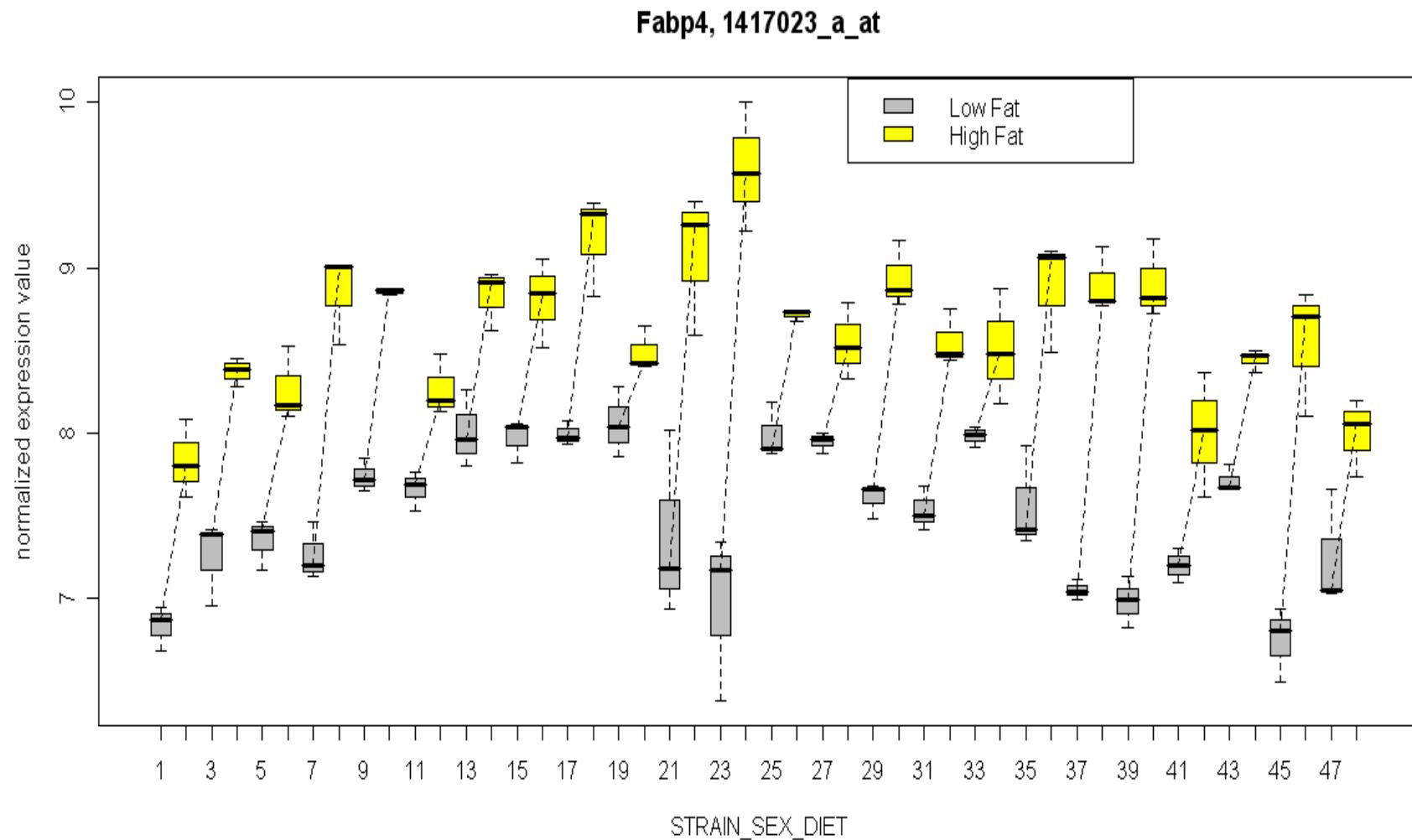


QC: Signal Intensity Data



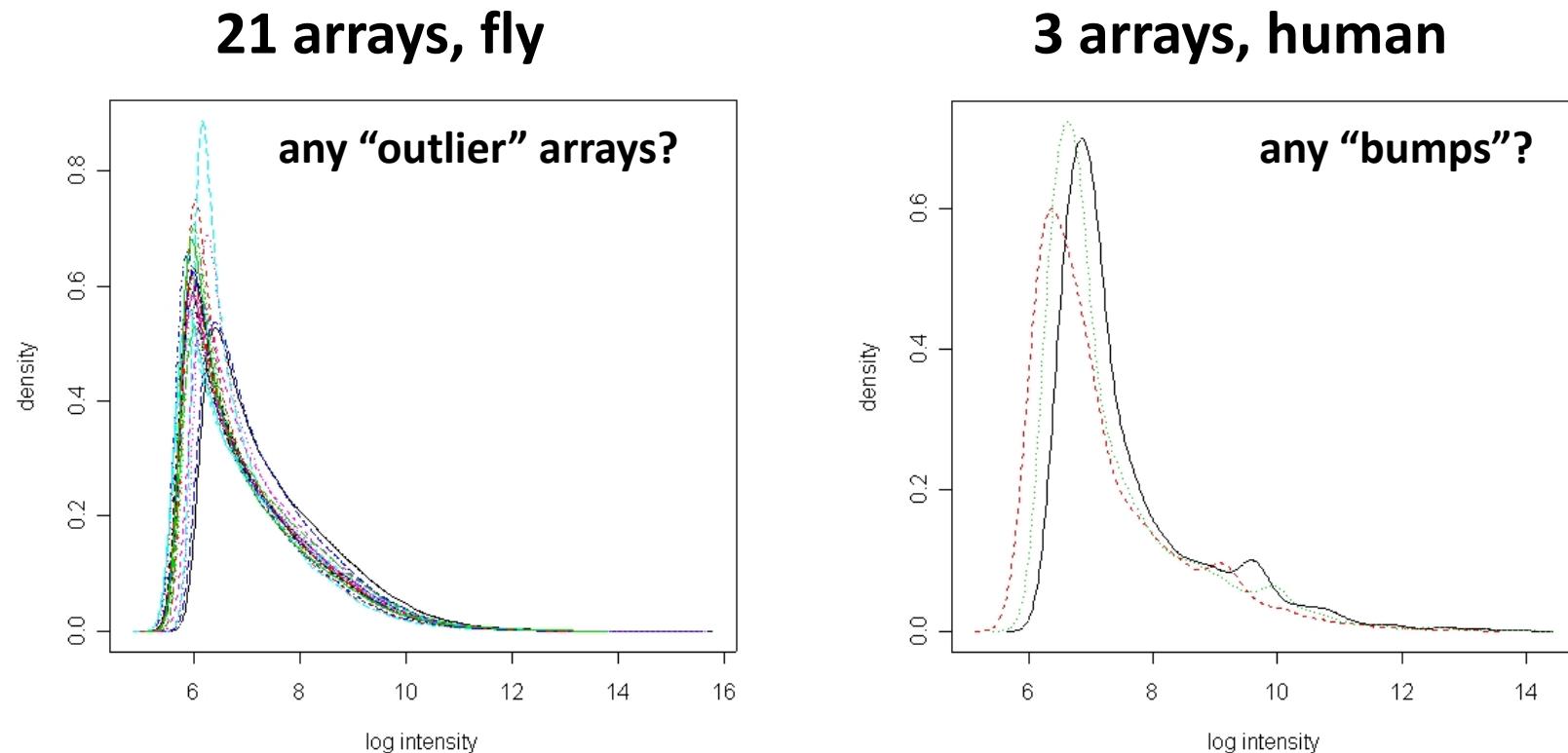
(R/affy library)

QC: Known Biological Response



(R/graphics library)

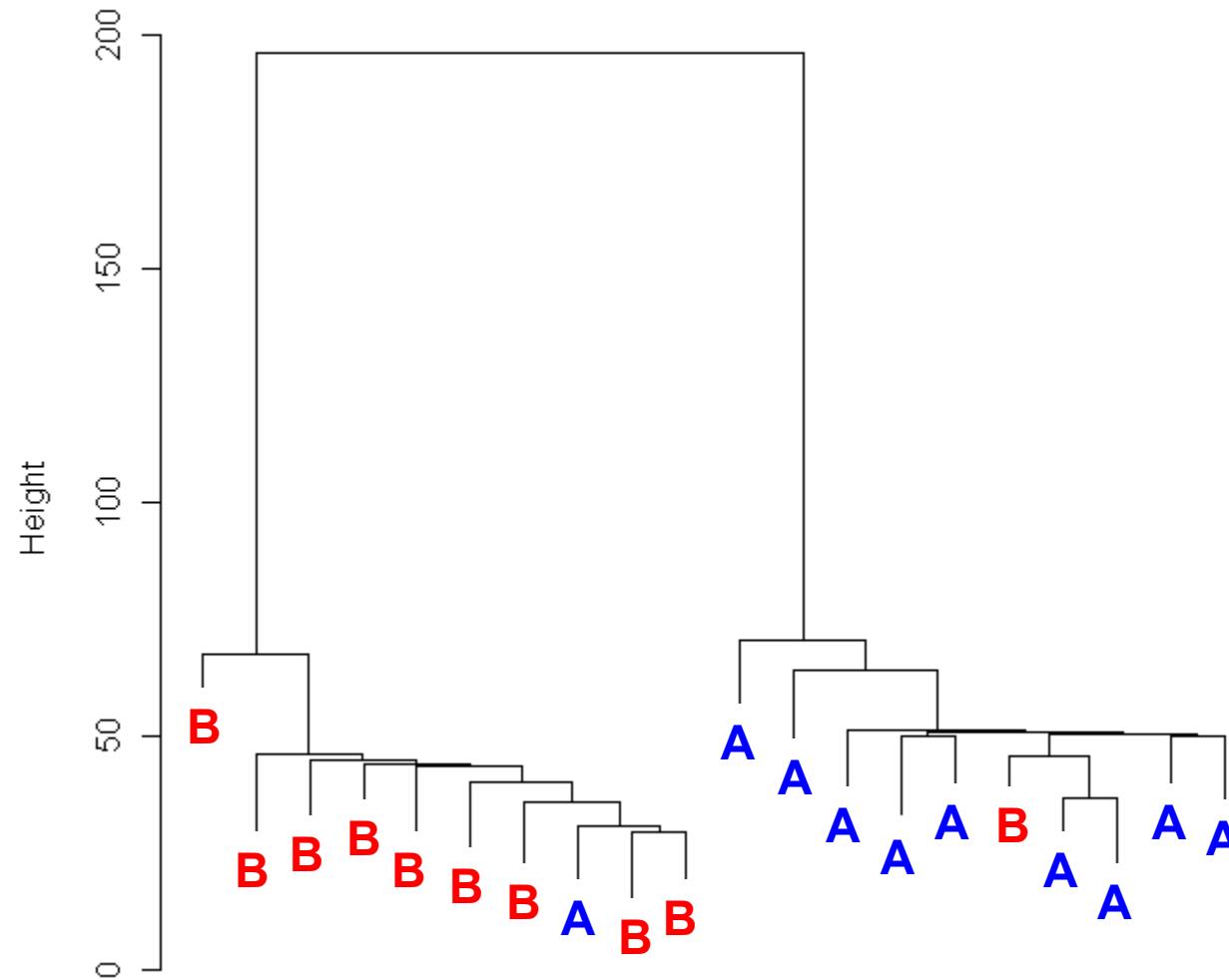
QC: Intensity Distributions



- Indicates need for normalization

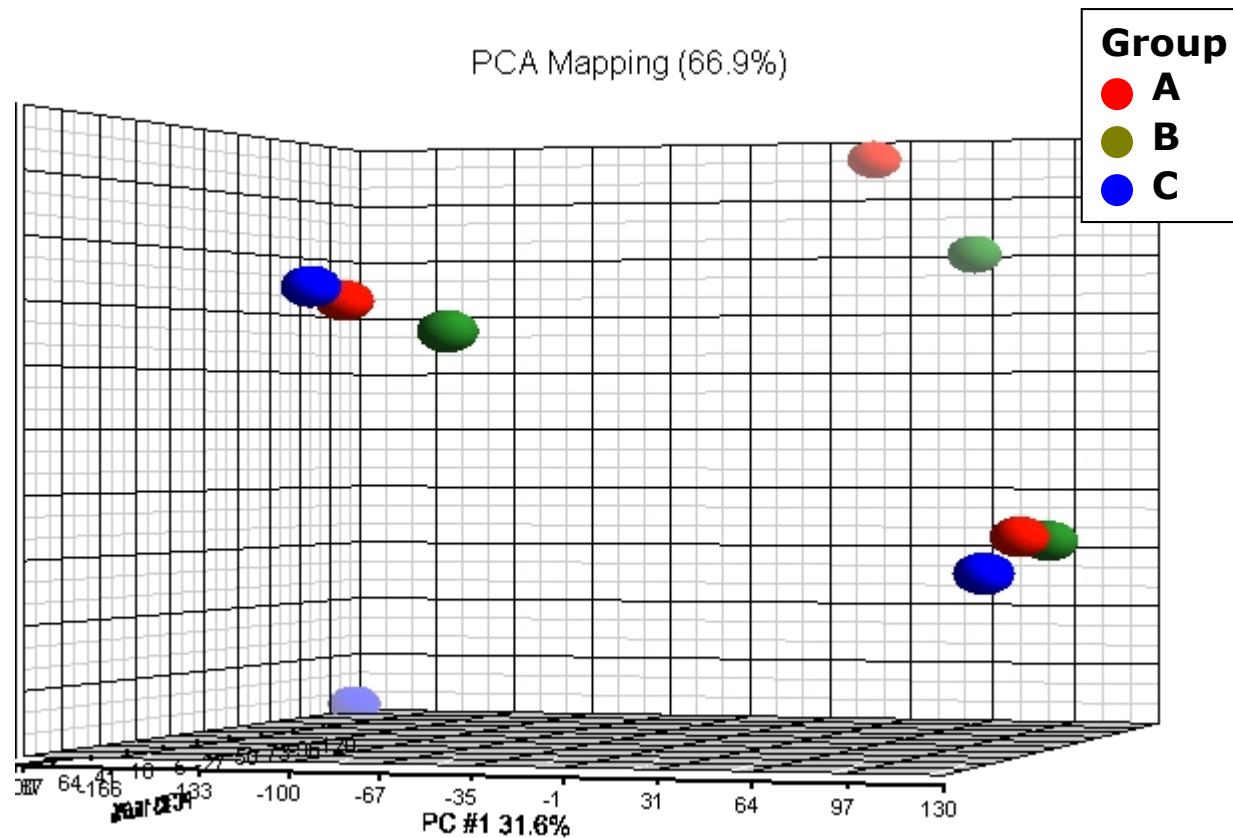
(R/graphics library)

QC: Cluster Analysis



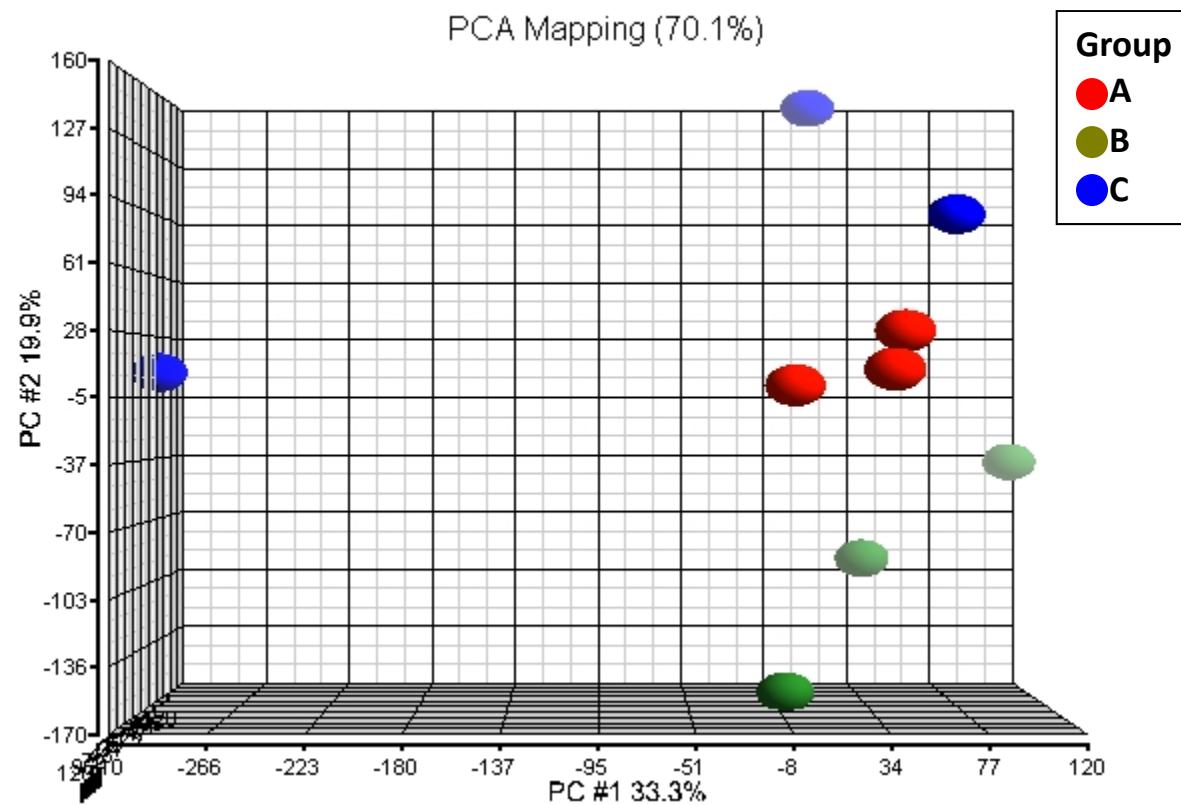
(R/graphics library)

QC: Batch Effects (uncorrected)



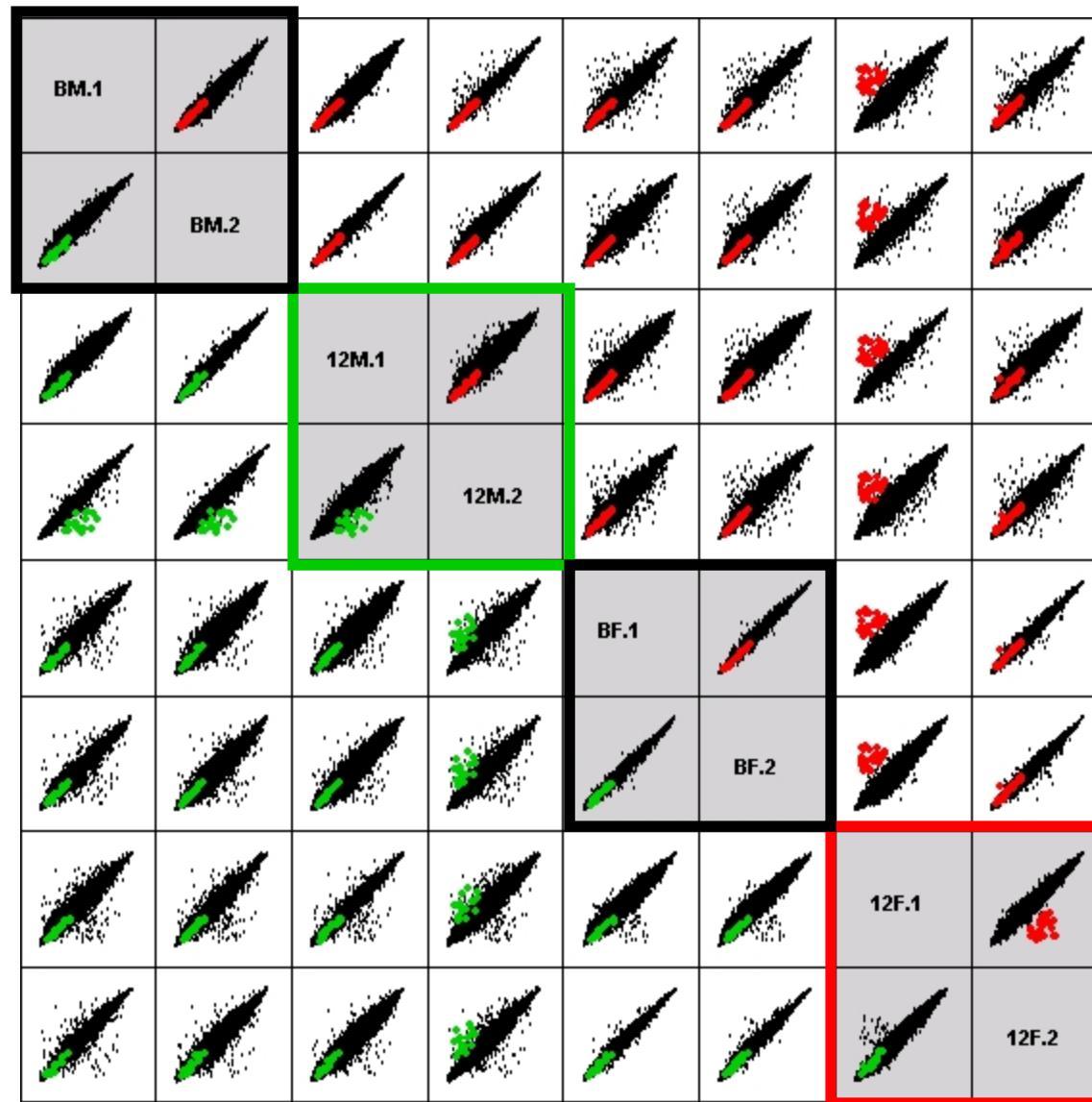
(Partek)

QC: Batch Effects (“corrected”)



(Partek)

QC: RMA Scatterplot



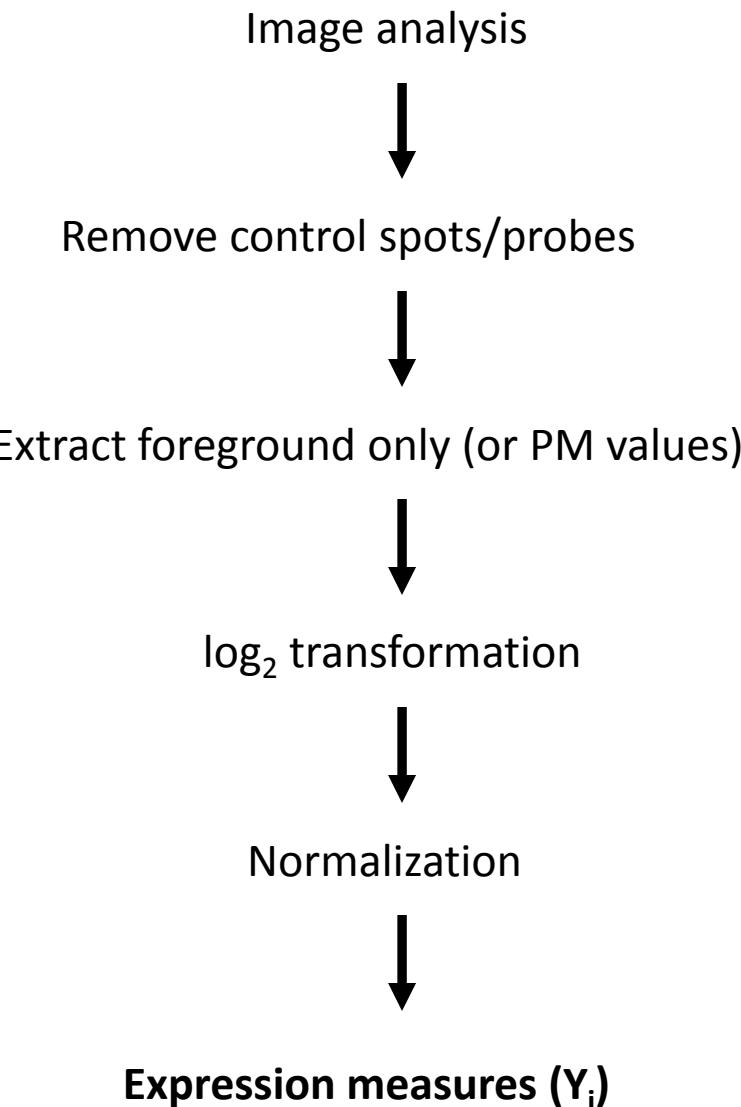
Muscle contraction,
Muscle development,
Hypoxia

Digestion of carbohydrates
and proteins

Shockley and Churchill, 2006

Data Preprocessing

Preprocessing

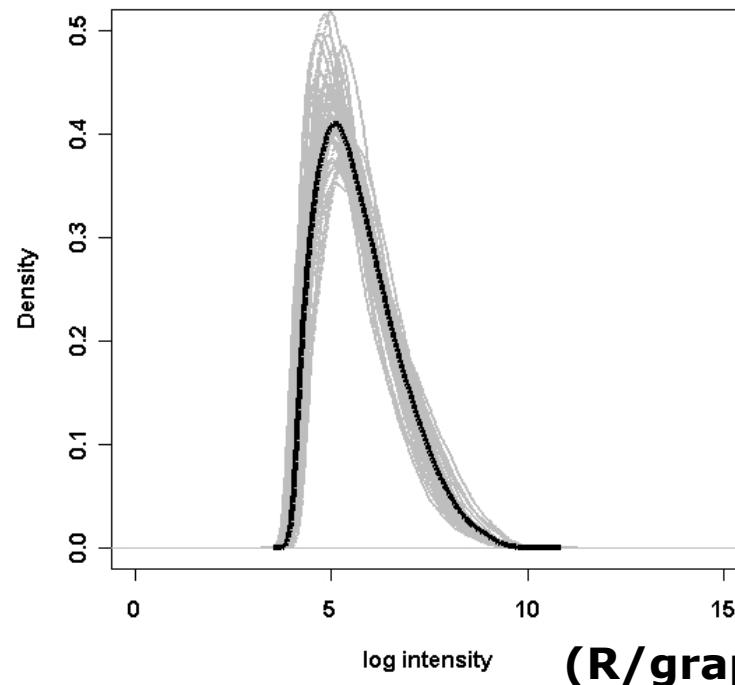


Expression Measures (Y_i)

I. Background Adjust PM values

~ 26% of MM values > PM values

II. Apply quantile normalization



(R/graphics library)

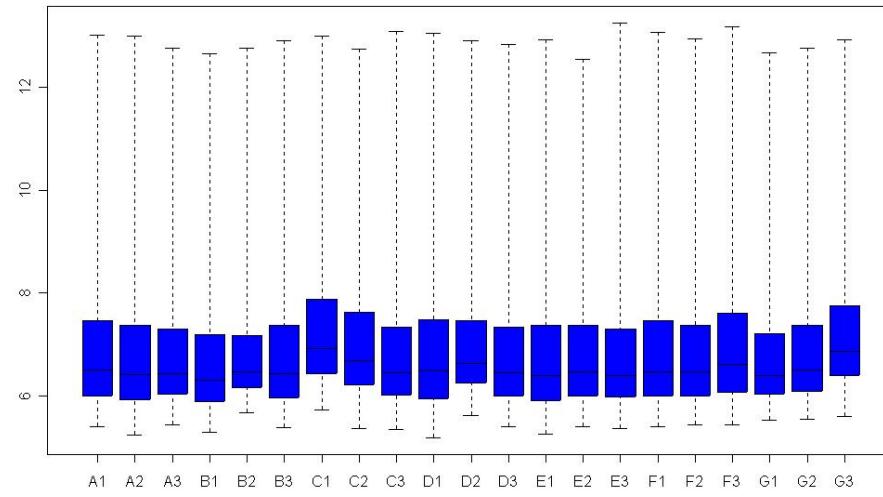
III. Summarize Probe Intensities

$$y_{ijn} = \mu_{in} + \alpha_{jn} + \varepsilon_{ijn}$$

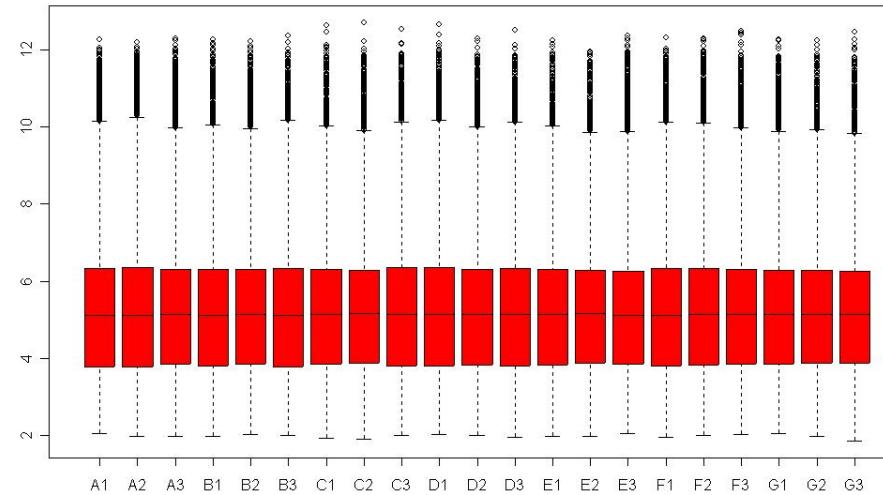
α_j – probe affinity effect; μ_{in} is expression measure

RMA (Irizarry et al., 2003)

Normalization



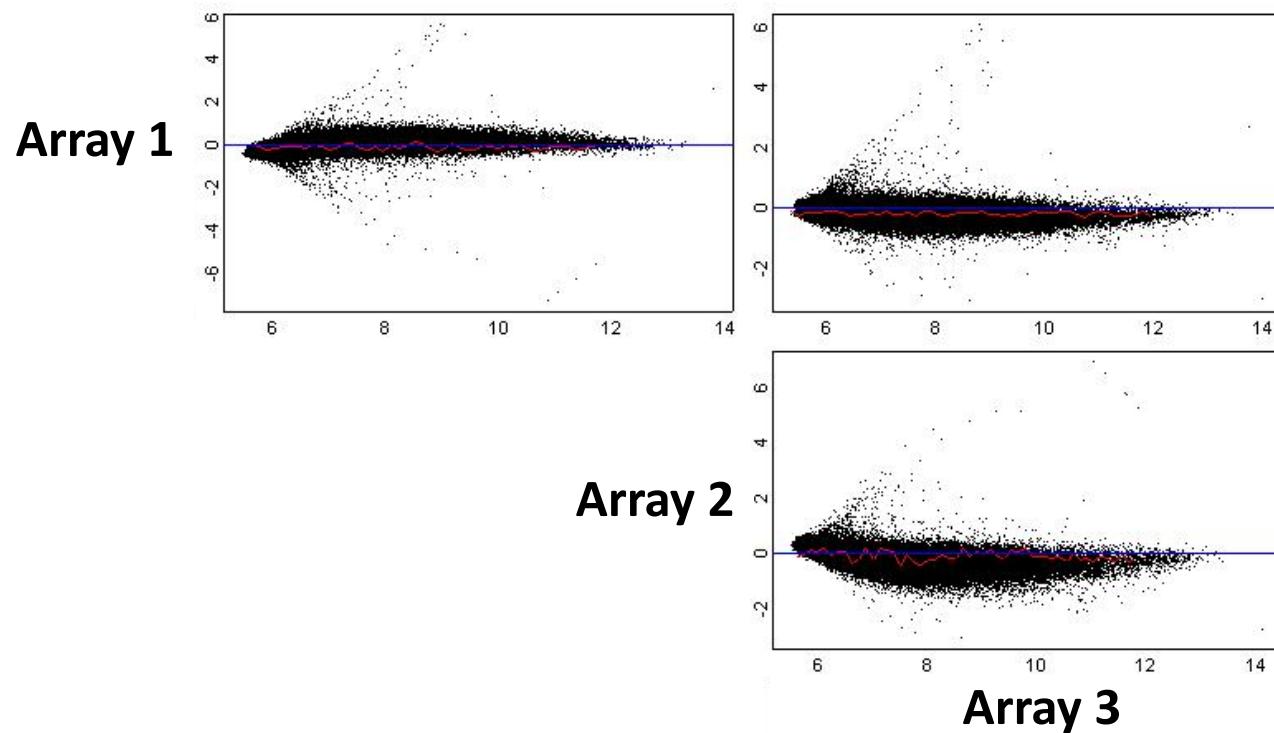
BEFORE



AFTER

(R/affy library)

MvA Plot



$$M = \log_2 A_i - \log_2 B_i = \log_2 (A_i / B_i)$$
$$A = (\log_2 A_i + \log_2 B_i)/2 = \log_2 \sqrt{A_i \times B_i}$$

(R/affy library)

Test for Differential Expression (ANOVA)

Null Hypothesis

“A hypothesis for which the effects of interest are assumed to be absent.”

Cui and Churchill, 2003

t-test

$$t = \frac{\bar{Y}_{g1} - \bar{Y}_{g2}}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right)}}$$

Y_{gi} : $\log_2(\text{expression measure condition } i)$

σ_i^2 : variance for condition i

n_i : sample size for condition i

F-test

$$F = \frac{\left(\frac{SS_{between}}{df_1} \right)}{\left(\frac{SS_{within}}{df_2} \right)}$$

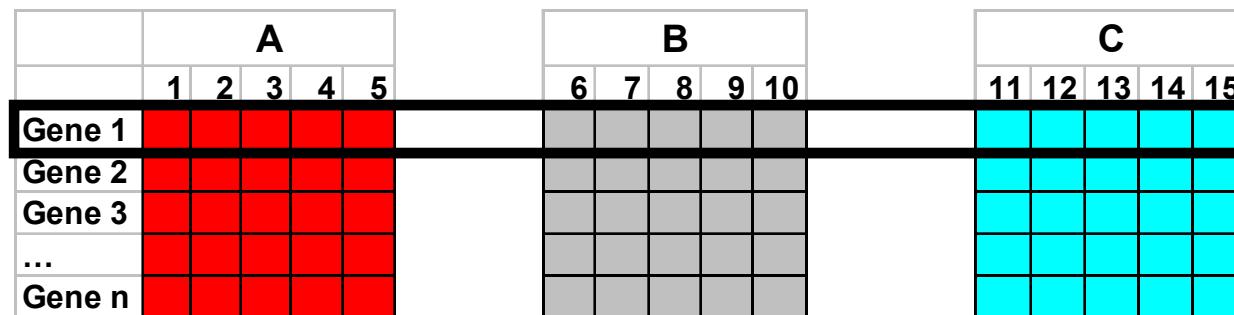
SS: the “sums of squares”,
a measure of variability

df1: number of groups – 1

df2: number of data – number of groups

ANOVA F-tests

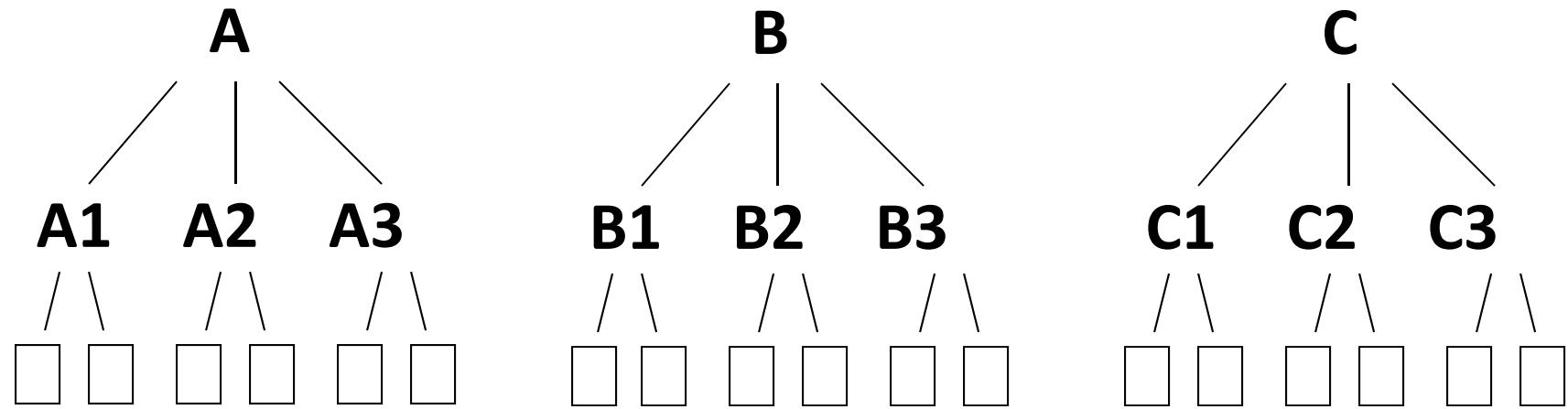
	Experiments														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Gene 1	Red	Red	Red	Red	Red	Grey	Grey	Grey	Grey	Grey	Cyan	Cyan	Cyan	Cyan	Cyan
Gene 2	Red	Red	Red	Red	Red	Grey	Grey	Grey	Grey	Grey	Cyan	Cyan	Cyan	Cyan	Cyan
Gene 3	Red	Red	Red	Red	Red	Grey	Grey	Grey	Grey	Grey	Cyan	Cyan	Cyan	Cyan	Cyan
...	Red	Red	Red	Red	Red	Grey	Grey	Grey	Grey	Grey	Cyan	Cyan	Cyan	Cyan	Cyan
Gene n	Red	Red	Red	Red	Red	Grey	Grey	Grey	Grey	Grey	Cyan	Cyan	Cyan	Cyan	Cyan



For Gene n, is the mean expression level the same across all time groups?

$$H_0: \mu_0 = \mu_1 = \mu_2$$

Example Affymetrix Data Set



A, B, C → mouse strains

ANOVA Model Development

$$Y_i = \mu + \text{STRAIN} + \sim\text{SAMPLE} + \varepsilon_i$$

↑
Test

Y_i – $\log_2(\text{expression measures})$

μ – gene mean (fixed)

STRAIN – factor (fixed)

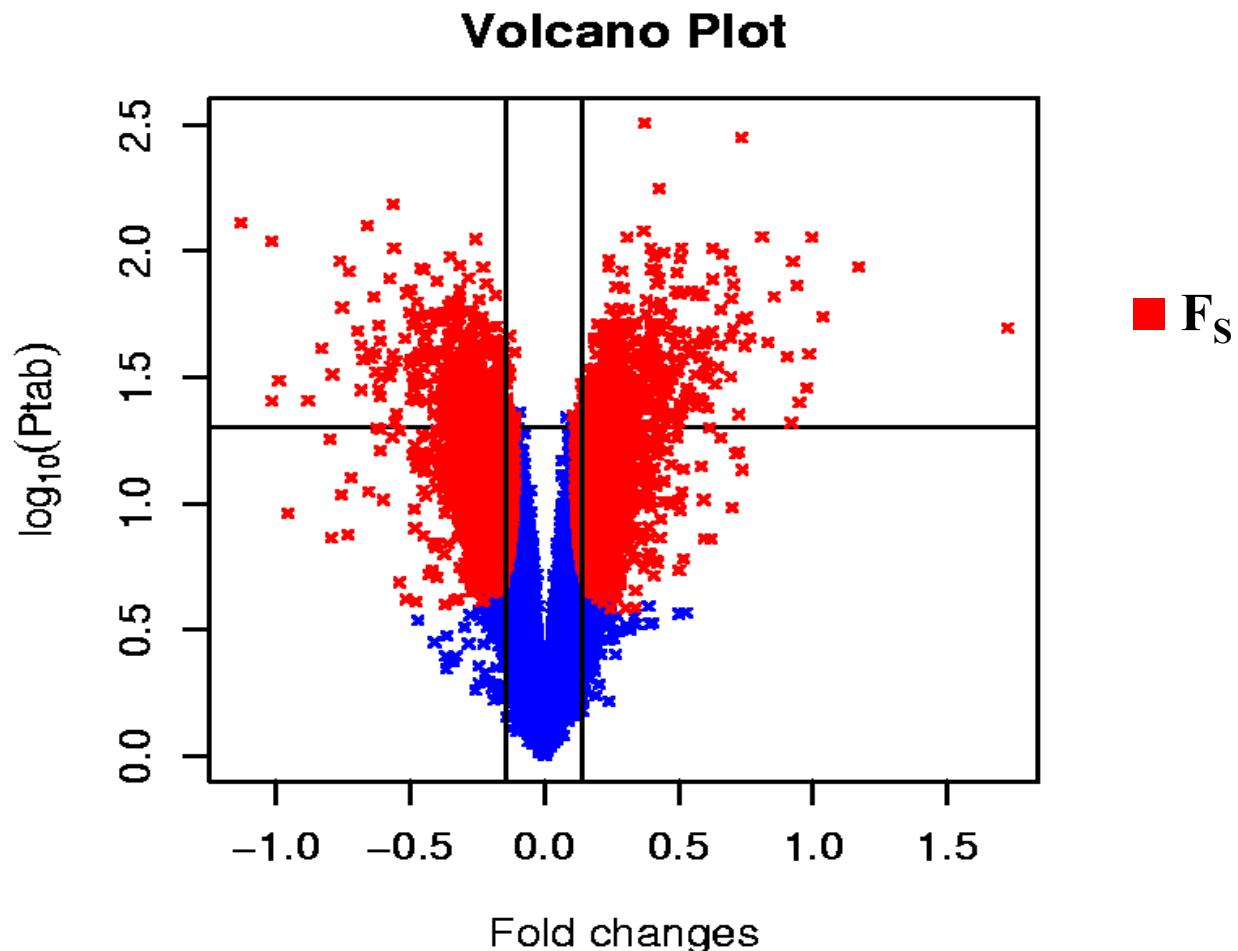
SAMPLE – factor (random)

ε_i – error (random)

Detecting Differential Expression

- One F statistic for each gene
 - F_1 gene specific variances (classical F-statistic)
 - F_s shrinks variances across genes (empirical Bayes)
- Determine significance (p-value):
 - Standard F-distribution table
 - Permutation testing
 1. compute F on all data
 2. permute subset data and compute F^*
 3. compare observed F to distribution of F^*

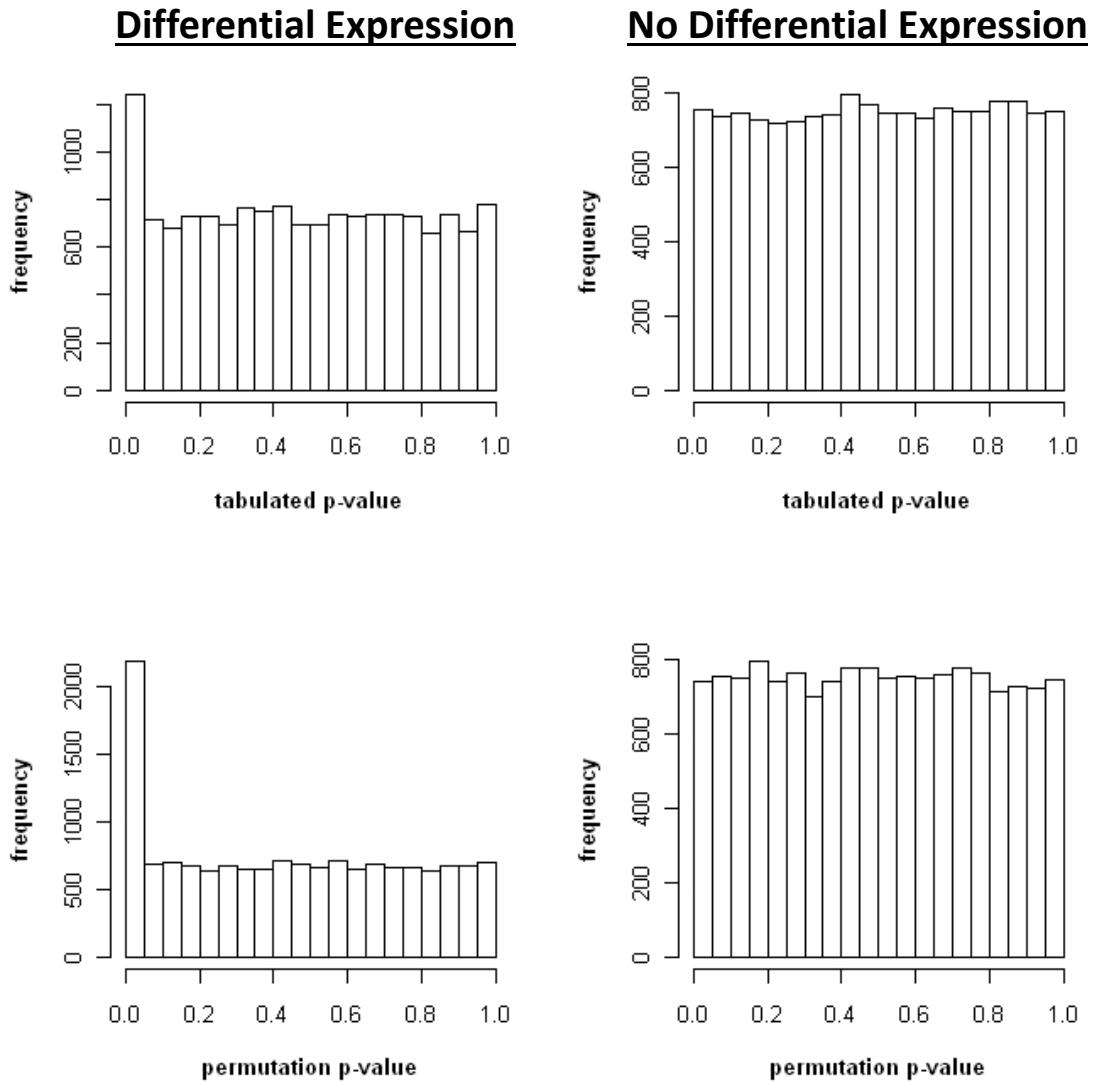
Modified F-statistic



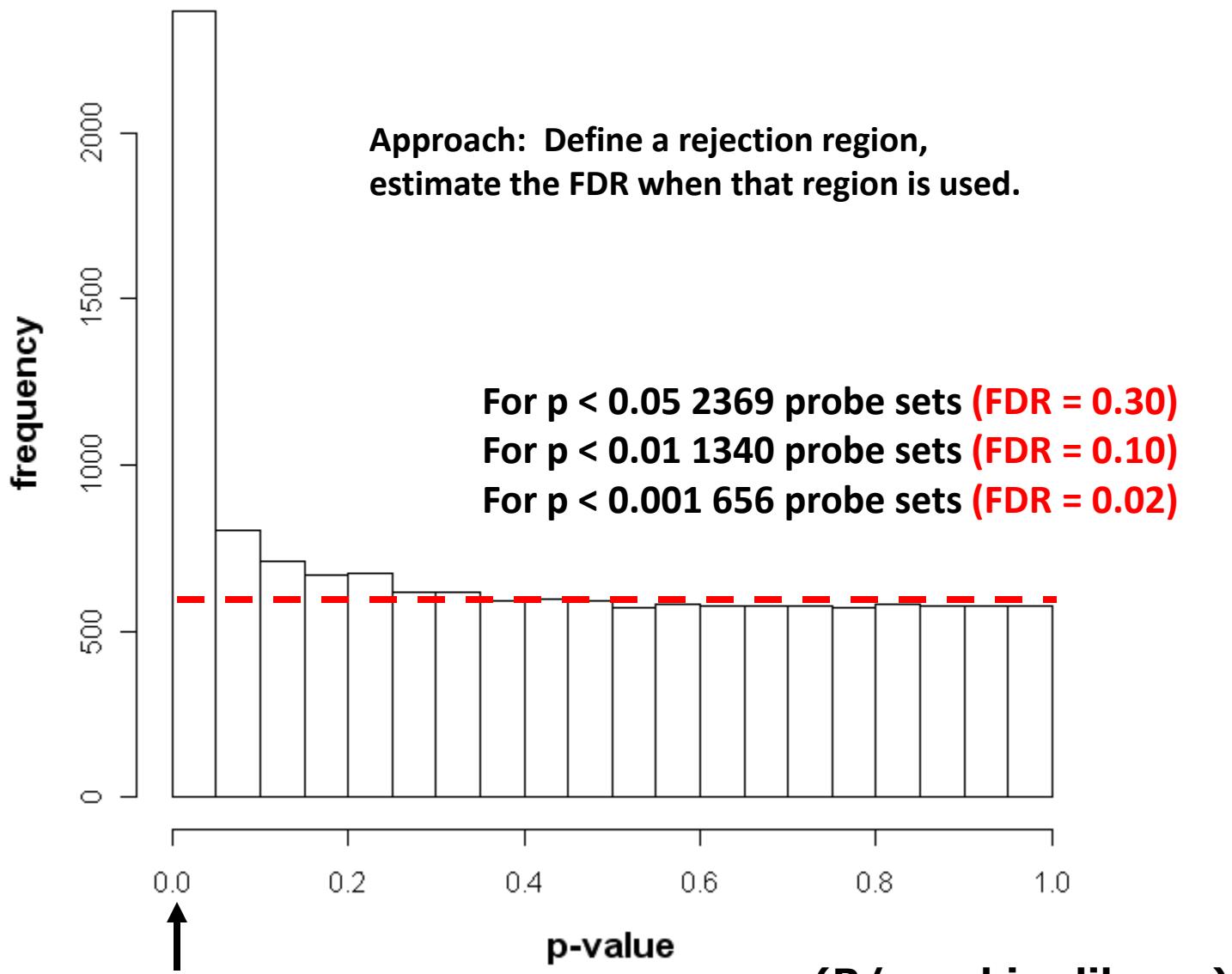
(R/maanova library)

source: pga.jax.org/hlb06coursefiles/Churchill_10-26.pdf

p-value distributions



False Discovery Rates

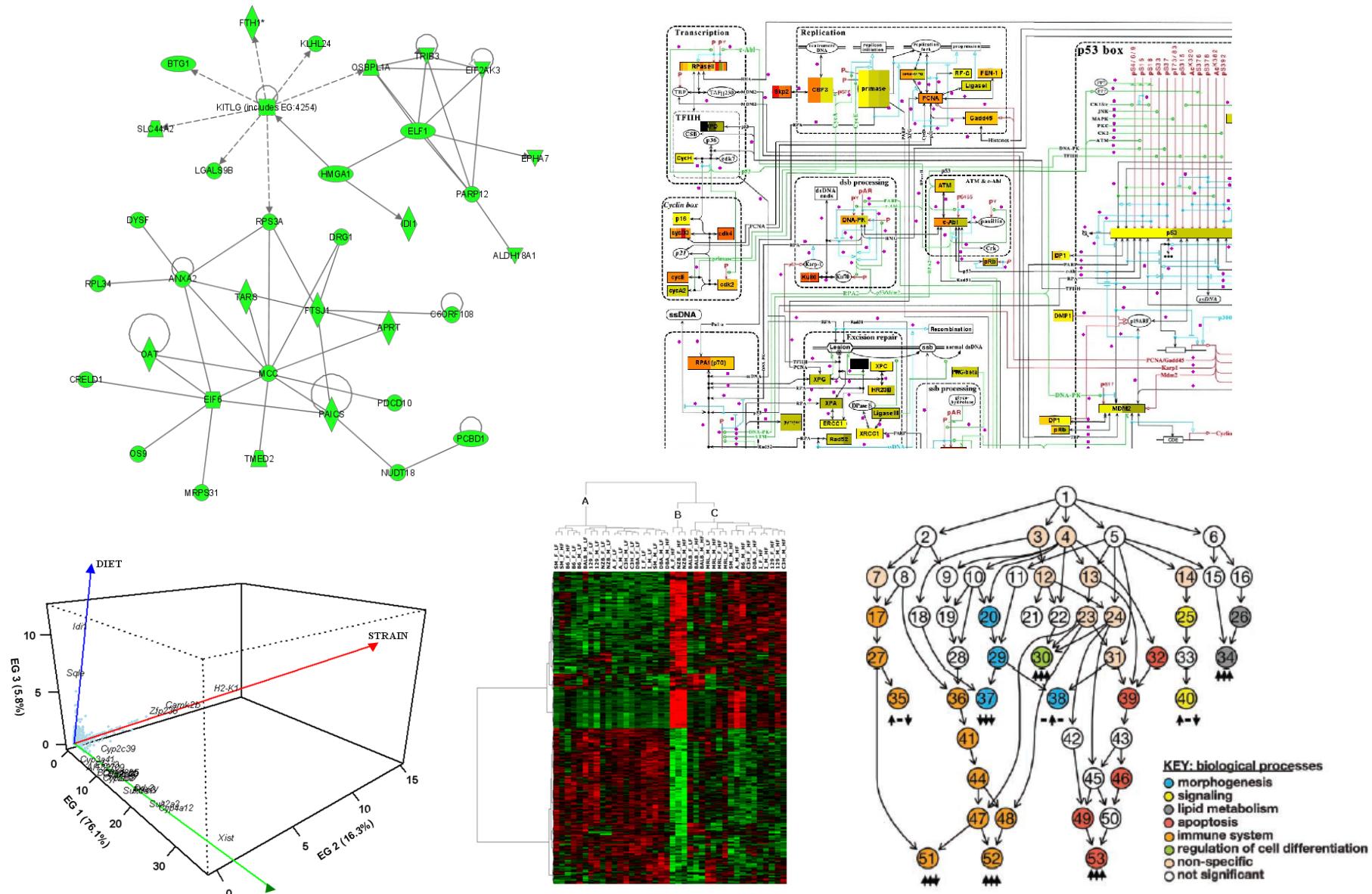


The Gene List

Probe Set ID	Gene Symbol	Entrez ID	Description	Fold Change	FDR	GO Terms	GO Term Descriptions
1455338_at	A4galt	239559	alpha 1,4-galactosyl	1.8	0.54	GO:0008150, GO:0008152	biological_process, metabolic
1420451_at	Accn5	58170	amiloride-sensitive c	1.5	0.71	GO:0008150, GO:0009987	biological_process, cellular prc
1416408_at	Acox1	11430	acyl-Coenzyme A o	2.3	0	GO:0008150, GO:0008152	biological_process, metabolic
1448318_at	Adfp	11520	adipose differentiatio	2.8	0	GO:0008150, GO:0008152	biological_process, metabolic
1422651_at	Adipoq	11450	adiponectin, C1Q ar	3.0	0	GO:0008150, GO:0065007	biological_process, biological r
1449019_at	Akap1	11640	A kinase (PRKA) ar	1.7	0		
1417130_s_at	Angptl4	57875	angiopoietin-like 4	6.5	0	GO:0008150, GO:0032502	biological_process, developme
1448839_at	Ankrd47	80880	ankyrin repeat doma	6.0	0.01		
1418849_x_at	Aqp7	11832	aquaporin 7	3.8	0	GO:0008150, GO:0009987	biological_process, cellular prc
1435108_at	Arhgap22	239027	Rho GTPase activat	1.3	0.57	GO:0008150, GO:0009987	biological_process, cellular prc
1445146_at	B230219N05Rik	319320	RIKEN cDNA B2302	1.4	0.76		
1434671_at	B230337E12Rik	98262	RIKEN cDNA B2303	1.4	0		
1452151_at	BC021523	223752	cDNA sequence BC	1.5	0.12	GO:0008150, GO:0008152	biological_process, metabolic
1434221_at	BC030863	194404	cDNA sequence BC	1.7	0.01	GO:0008150, GO:0065007	biological_process, biological r
1429539_at	Bcl2l13	94044	BCL2-like 13 (apopt	1.6	0	GO:0008150, GO:0032502	biological_process, developme

- Gene annotation, statistical summaries, etc...
- Statistical thresholds are arbitrary, but good place to begin

Gene Lists Are Starting Points!



R: Search Pubmed Abstracts

```
> library("mouse4302.db"); library("annotate"); library("XML");
>
> probeNames <- data$ProbeID[1:10]
> probeNames
[1] "1415670_at"    "1415671_at"    "1415672_at"    "1415673_at"
[6] "1415675_at"    "1415676_a_at"   "1415677_at"    "1415678_at"   "1415674_a_at"
>                                         "1415679_at"
>
> absts <- pm.getabst(probeNames, "mouse4302.db");
Read 6820 items
Read 6875 items
Read 5822 items
Read 6468 items
Read 4583 items
Read 4897 items
Read 6331 items
Read 6262 items
Read 7229 items
Read 5302 items
> absts[[1]][[1]]
An object of class 'pubMedAbst':
Title: Normalization and subtraction: two approaches to facilitate gene
discovery.
PMID: 8889548
Authors: MF Bonaldo, G Lennon, MB Soares
Journal: Genome Res
Date: Sep 1996
```

R/annotate

R: Search Pubmed Abstracts

```
> titl <- sapply(absts[[1]], articleTitle);
> titl[1]
[1] "Normalization and subtraction: two approaches to facilitate gene discovery."
>
> myfunc <- function(x) pm.abstGrep("[Pp]rotein", x)
> pro.res <- sapply(absts, myfunc)
> pro.res[[2]]
[1] TRUE FALSE TRUE FALSE FALSE FALSE TRUE FALSE FALSE FALSE TRUE TRUE
[13] TRUE FALSE TRUE TRUE
> pmAbst2HTML(absts[[2]], filename="pm.html")
```

BioConductor Abstract List

Article Title	Publication Date
Brain Ac39/physophilin: cloning, coexpression and colocalization with synaptophysin.	Mar 1998
High-efficiency full-length cDNA cloning.	Month 1999
Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray.	Aug 2000
Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes.	Oct 2000

R/annotate

Clustering

Microarray Gene Clustering

- Idea: Genes with similar function will have similar expression patterns in microarray experiments
- Two main types:
 - Hierarchical (nested)
 - Partitional (k -means)
- Search the data to find groups of genes (**clusters**) based on a measure of similarity or dissimilarity (**distance metric**)
- A “cluster” refers to the set of genes that are alike

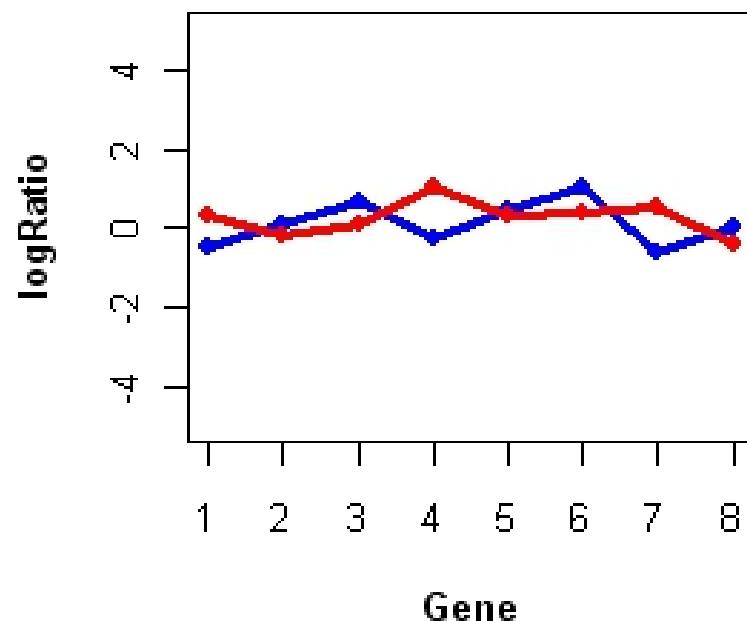
Similarity or Dissimilarity (1)

- Distance function measures how different two expression patterns are
- Distance properties:
 - $d(x,x) = 0$
 - $d(x,y) > 0$
 - $d(x,y) = d(y,x)$
- Common distance metrics:
 - Euclidean
 - Manhattan
 - Correlation

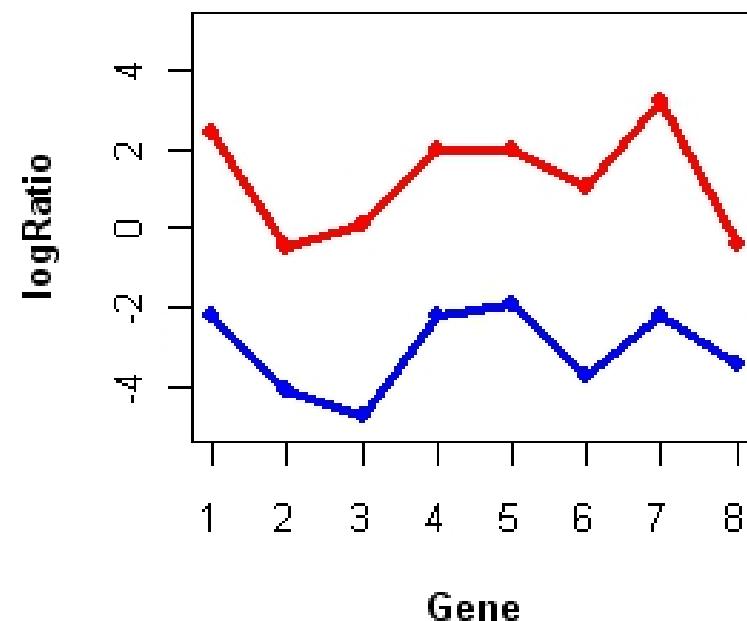
Similarity or Dissimilarity (2)

█ treated
█ untreated

Euclidean



Correlation



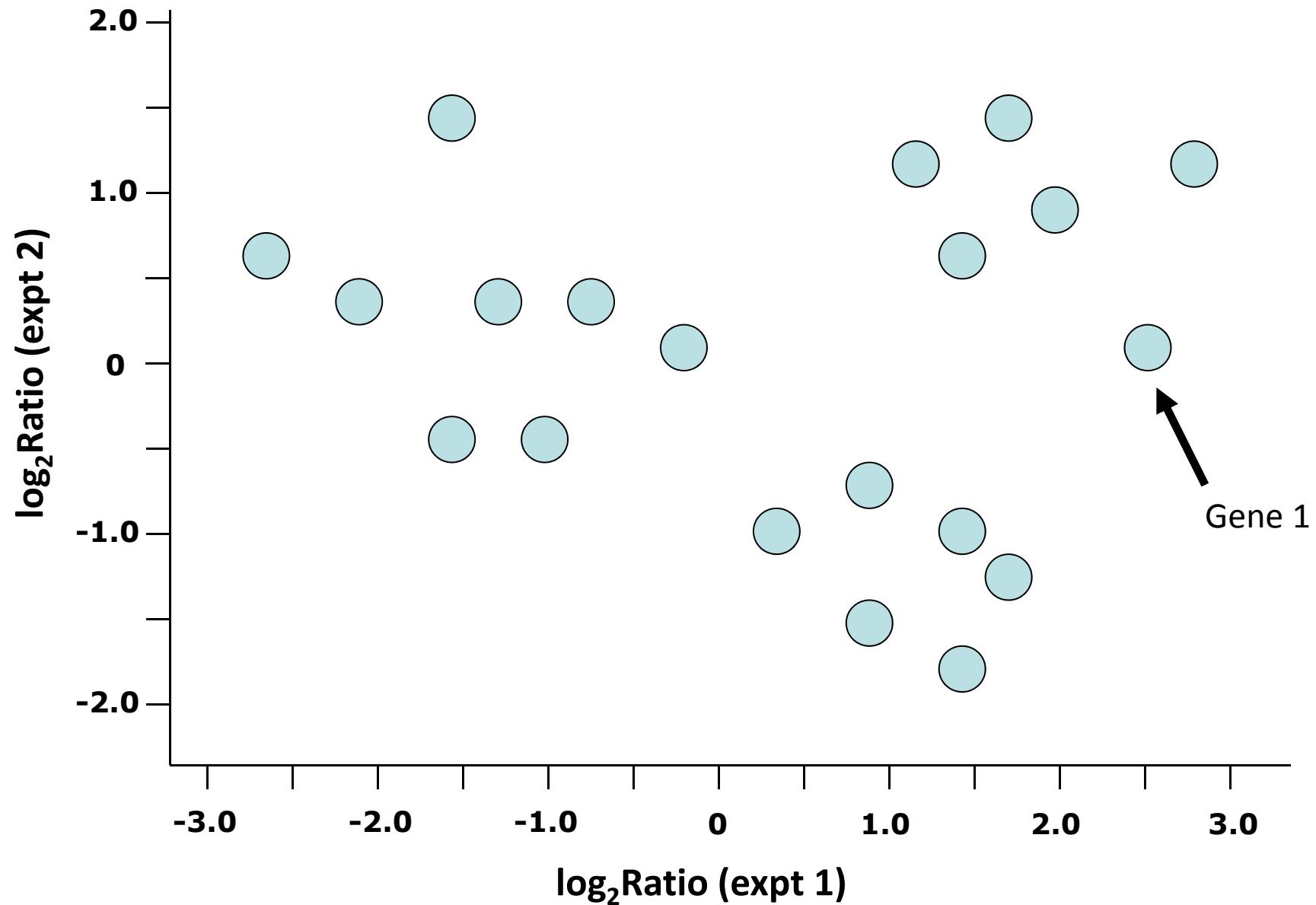
Small Euclidean distance
Large Corr distance

Large Euclidean distance
Small Corr distance

k -means Cluster Algorithm

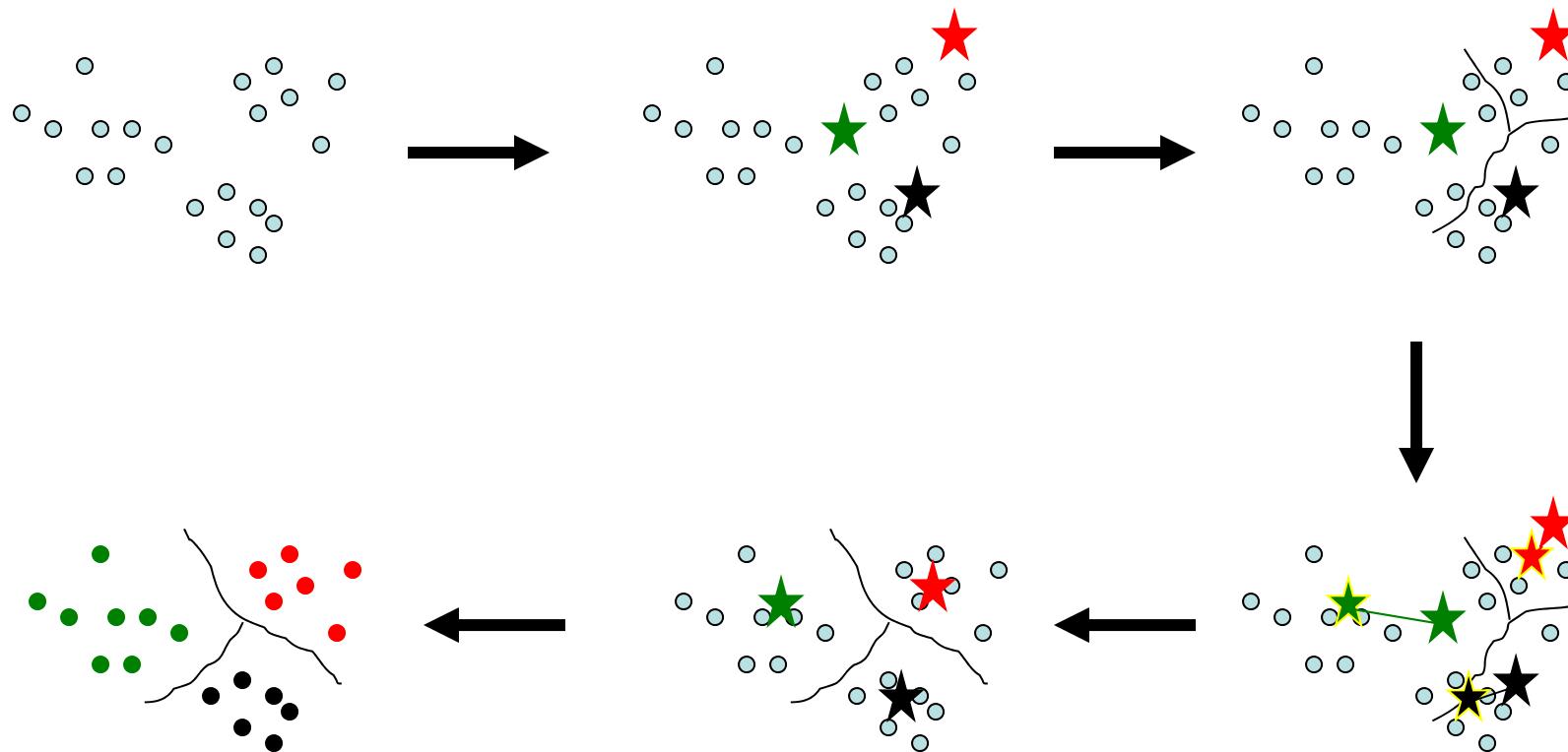
1. Place k points into the space of all objects (the k points represent the first “centroids”)
2. Assign each object to the group with the closest centroid
3. Recalculate positions of all centroids
4. Repeat (2) and (3) until centroids no longer move

k -means Cluster Algorithm

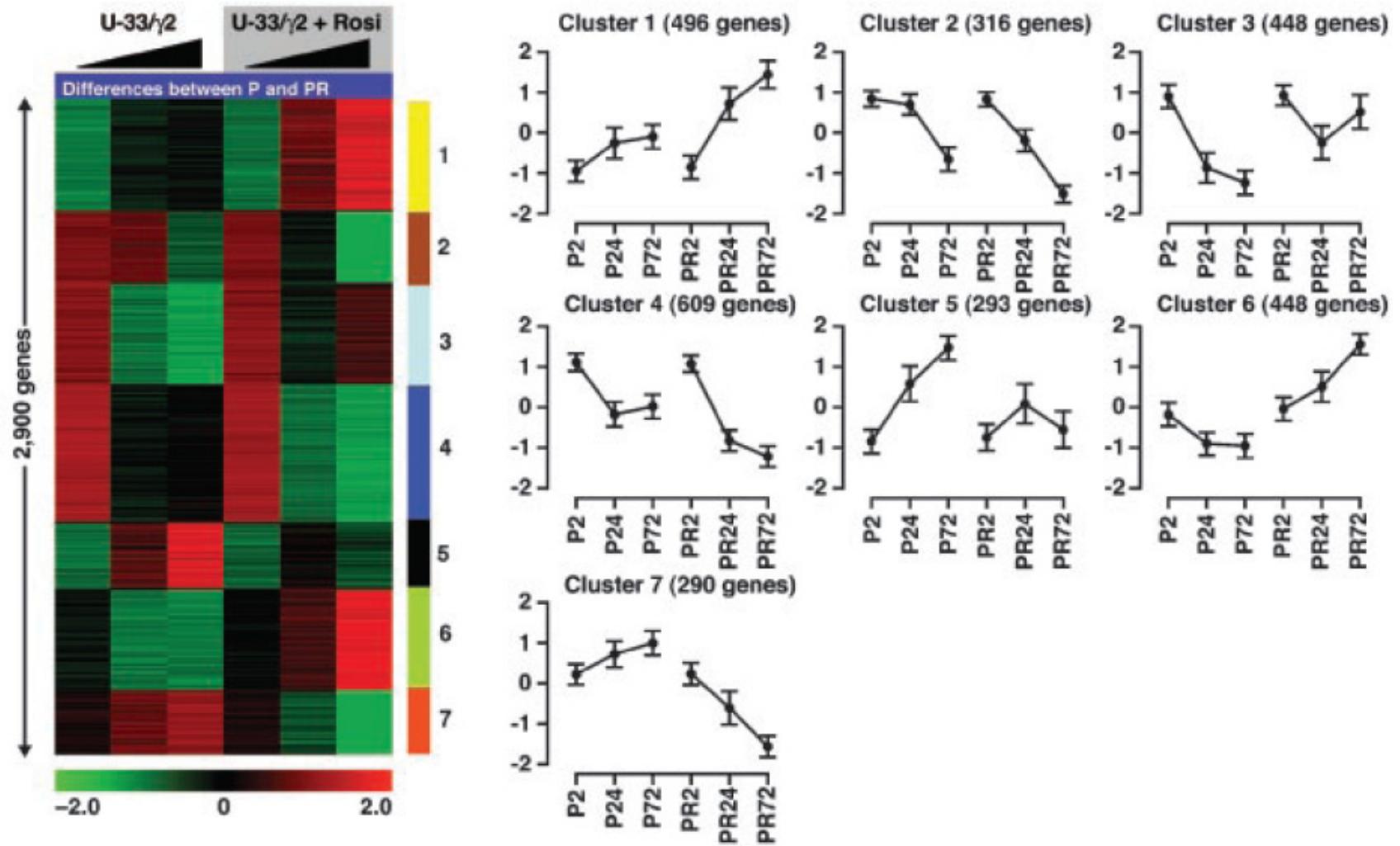


k -means Cluster Algorithm

Try $k = 3$



Example k -means Cluster

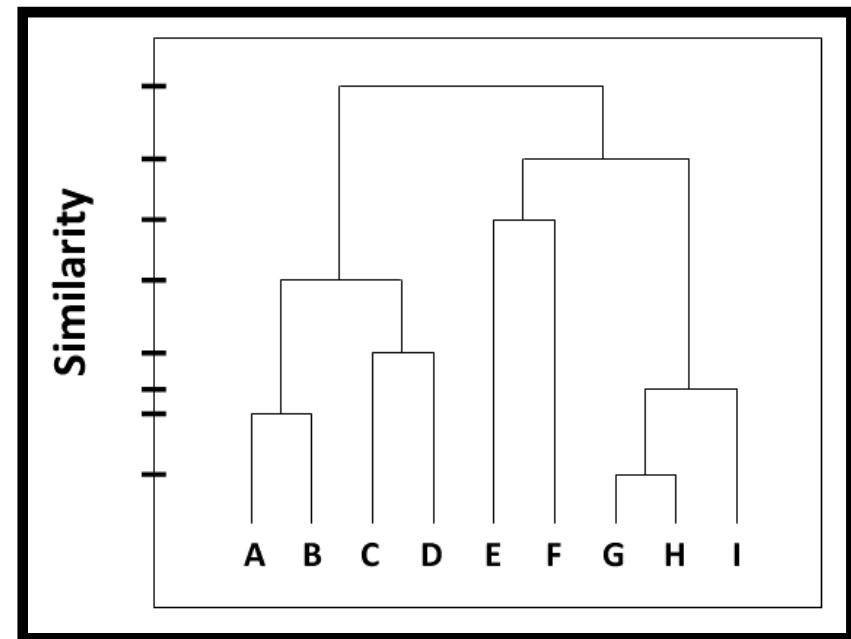
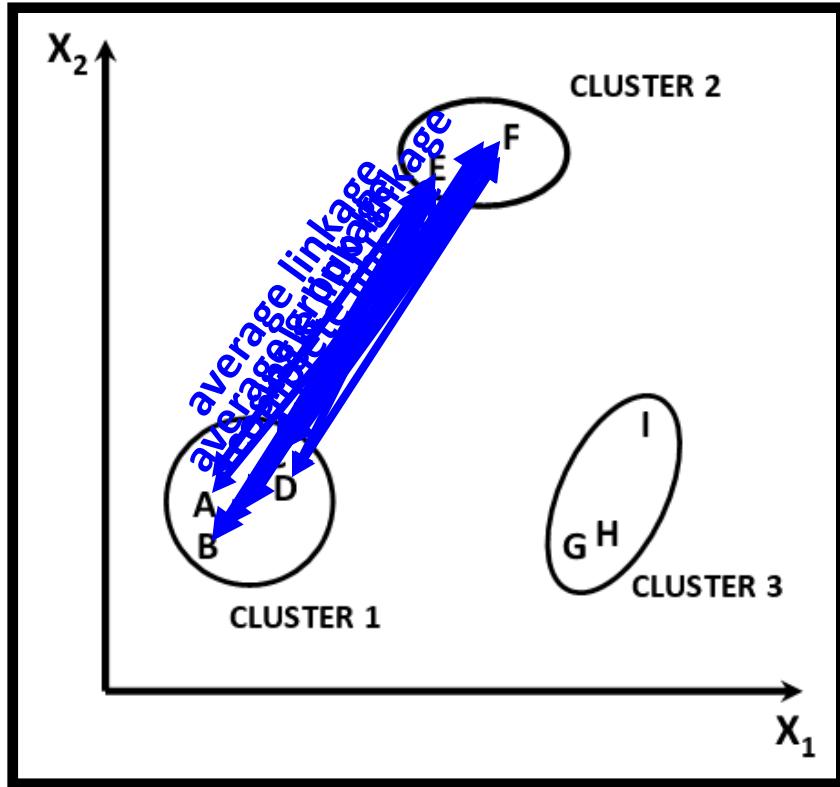


Shockley et al, 2009a

Hierarchical Cluster Algorithm

1. Calculate the pairwise distance matrix between all genes
2. Start by considering each gene by itself as a separate cluster
3. The two “clusters” with the smallest distance are merged to form a single cluster
4. Recalculate the pairwise distance matrix between the remaining genes and the new cluster
5. Repeat (3) and (4) until only one cluster remains

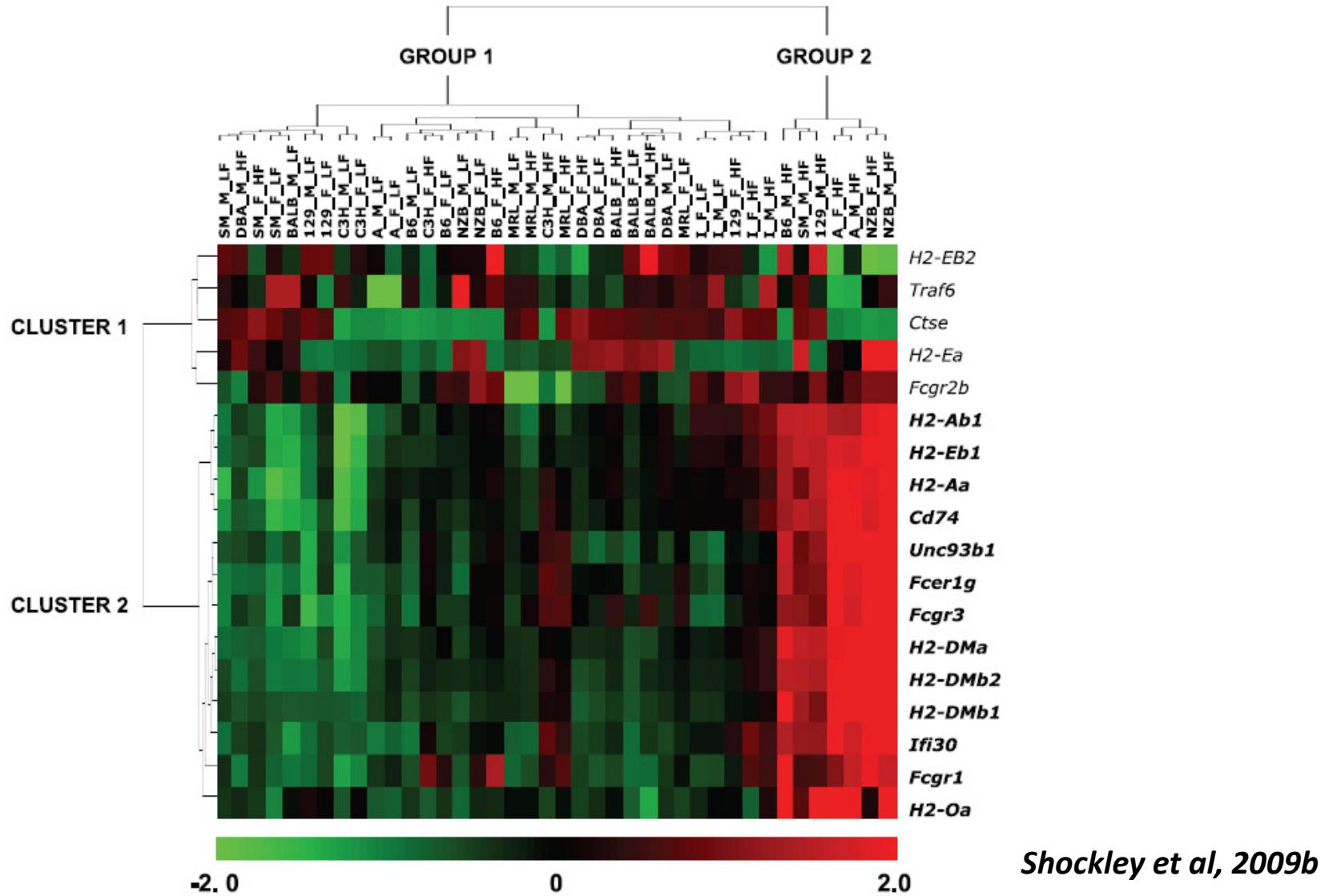
Hierarchical Clustering



Linkage Method

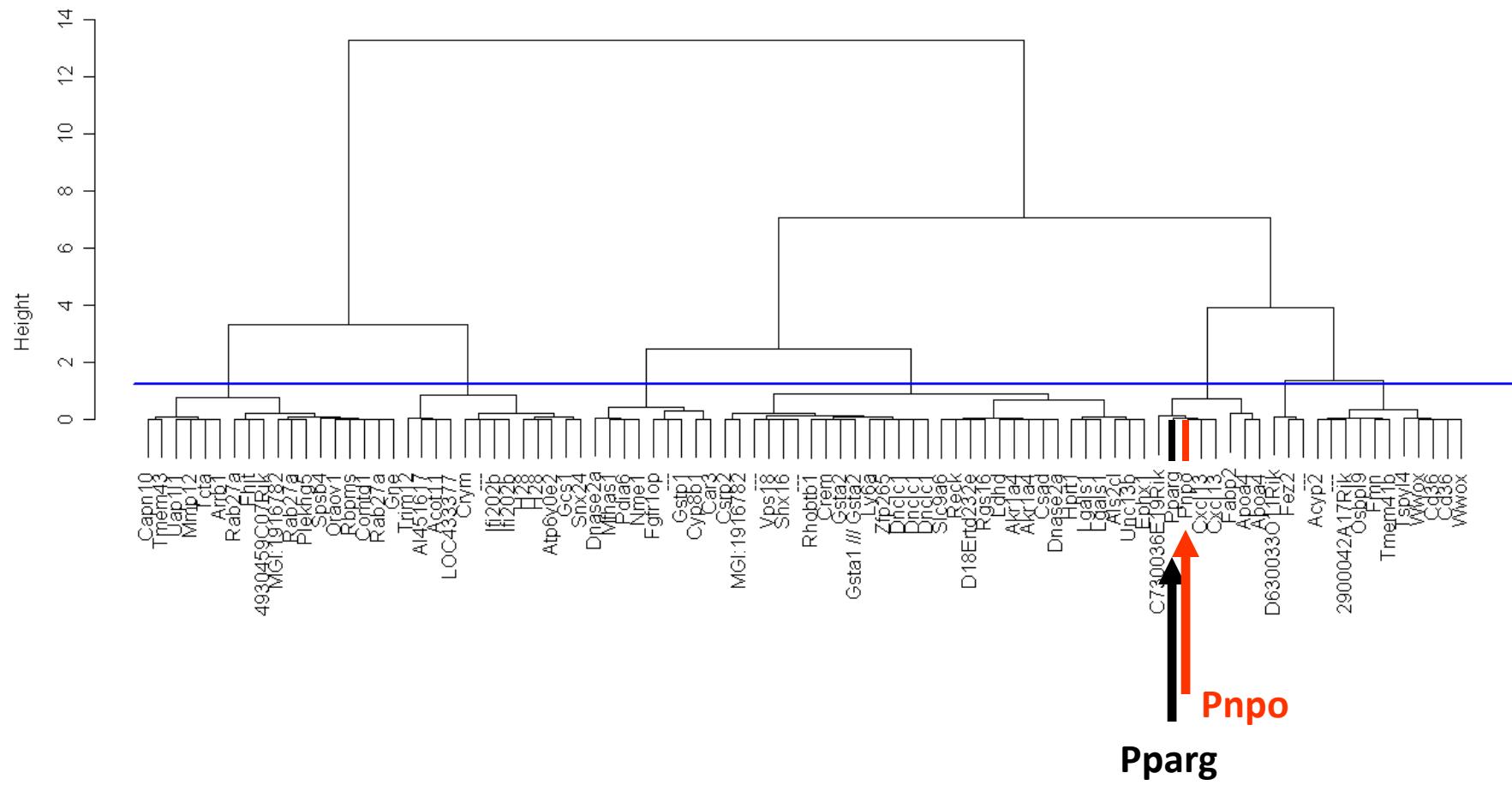
1. Single linkage
2. Complete linkage
3. Average linkage
4. Average group linkage
5. Ward's Method (minimize error sum of squares)

Example Hierarchical Cluster



Hierarchical Clustering

Pnpo - pyridoxamine-phosphate oxidase activity
Vitamin B6 metabolism [Chromosome 11]



(R/graphics)

Clustering

- Microarray data is typically filtering and normalized before applying cluster analysis
- Attempt to find natural groups in the data
- Useful for data reduction and hypothesis generation
- After finding clusters, identify all genes and look for a common function
- Heat maps are useful for visualization and data summarization

Gene Expression Omnibus

The screenshot shows the GEO homepage integrated with NCBI. At the top left is the NCBI logo. The top right features the GEO logo and navigation links for GEO Publications, FAQ, MIAME, and Email GEO. A user status message "Not logged in | Login" is also present. Below the header, a main content area includes:

- GEO navigation:** A tree diagram under the "QUERY" section branches into DataSets, Gene profiles, GEO accession, and GEO BLAST, each with a "GO" button. Under the "BROWSE" section, it branches into DataSets, Platforms, Samples, and Series, with GEO accessions connecting to the latter three.
- Site contents:** A sidebar listing various resources:
 - Public data: Platforms (7,502), Samples (449,490), Series (17,523).
 - Documentation: Overview, FAQ, Find, Submission guide, Linking & citing, Journal citations, Programmatic access, DataSet clusters, GEO announce list, Data disclaimer, GEO staff.
 - Query & Browse: Repository browser, Submitters, SAGEmap, FTP site, GEO Profiles, GEO DataSets, Submit, New account.
- Submitter login:** Fields for User id and Password, and buttons for LOGIN, » New account, and » Recover password.

At the bottom, a footer bar contains links for NLM, NIH, Email GEO, Disclaimer, and Section 508.

MIAME:

1. Raw Data
2. Normalized Data
3. Sample Annotation
4. Experimental Design
5. Array Annotation
6. Laboratory and Data Processing Protocols

source: <http://www.ncbi.nlm.nih.gov/geo/info/MIAME.html>

Summary

- Experimental design should be determined by scientific question
- Biological replicates are required to account for normal variation
- Use RNA of the highest quality
- Employ good quality and process control practices
- Robust statistical methods needed to distinguish signal from noise
- Always validate!



Useful References

- Ayroles JF, Gibson R. (2006). Analysis of variance of microarray data. *Methods in Enzymology.* 411:214-233.
- Brazma et al (2001). Minimum information about a microarray experiment (MIAME). *Nature Genetics.* 29: 365-371.
- Churchill GA. (2002). Fundamentals of experimental design for cDNA microarrays. *Nature Genetics.* 32 Suppl:490-495.
- Churchill GA. (2004). Using ANOVA to analyze microarray data. *Biotechniques.* 37: 173-175.
- D'haeseleer P. (2005). "How does gene expression clustering work?" *Nature Biotechnology* 23: 1499-1501.
- Gibson G. (2003). Microarray analysis: genome-scale hypothesis scanning. *PLoS* 1:E15.
- Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP (2003b). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research* 31: e15.
- Jain AK, Murty MN, Flynn PJ. (1999). Data Clustering: A review. *ACM Computing Surveys.* 31: 264-323.
- Quackenbush, J. (2002). Microarray data normalization and transformation. *32 Suppl:* 496-501.
- Schena M, Shalon D, Davis RW, Brown PO. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science.* 270: 467-470.

Useful R Packages

■ Preprocessing/Normalization (Bioconductor Repository)

<http://www.bioconductor.org/packages/release/Software.html>

affy – assessment of affymetrix data

affycomp – assessment of expression measures

affypdnn – probe dependent nearest neighbor

affyPLM – probe level models

annaffy – annotation tools for biological metadata

makecdfenv – CDF environment maker

simpleaffy – affy based high level functions

vsn – variance stabilization and calibration

■ Affy Data Sets

affydata - <http://www.bioconductor.org/packages/devel/data/experiment/html/affydata.html>

estrogen - <http://www.bioconductor.org/data/experimental.html>

spike-in data - <http://affycomp.biostat.jhsph.edu/>

■ Microarray Data Analysis Packages

ebarays - <http://www.biostat.wisc.edu/~kendzior/>

limma - <http://bioinf.wehi.edu.au/limma/>

maanova - <http://churchill.jax.org/software/rmaanova.shtml>

samr - <http://www-stat.stanford.edu/~tibs/SAM/Rdist/index.html>

Other Useful Software

- EDGE - Extraction of Differential Gene Expression (open source)
<http://www.genomine.org/edge/>
- EPIG – Extracting microarray gene Expression Patterns and Identifying co-expressed Genes (open source)
<http://www.niehs.nih.gov/research/resources/software/epig/index.cfm>
- J/maanova – Java tool for MicroArray ANalysis Of Variance (open source)
<http://churchill.jax.org/software/jmaanova.shtml>
- JMP® Genomics 4 (commercial)
<http://www.jmp.com/software/genomics/>
- Partek Genomics Suite (commercial)
<http://www.partek.com/>
- SAM – Significance Analysis of Microarrays (open source)
<http://www-stat.stanford.edu/~tibs/SAM/>
- ORIGEN – Order Restricted Inference for Ordered Gene ExpressioN (open source)
<http://www.niehs.nih.gov/research/resources/software/oriogen/index.cfm>
- NIEHS Tools!
<http://www.niehs.nih.gov/research/resources/software/>

THANK
YOU



Thanks for your attention

